Positional proteomics: advanced strategies for targeted proteome simplification

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

by

Lucy McDonald

October 2008

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Professor Rob Beynon, for his guidance and support throughout my time in Liverpool. I would like to extend this thanks to all members of the Proteomics and Functional Genomics Group for their help, support and friendship over the years. Special thanks to Lynn McLean, Duncan Robertson and Deborah Simpson for specialist training and advice.

The work in this thesis has been carried out over two research grants. The first grant, in which the initial N-terminal method was developed, was funded by NERC and was in collaboration with Jane Hurst, Malcolm Bennett and Paula Stockley. The second grant was funded by EPSRC and involved enhancement of the original N-terminal protocol along with the human plasma and the MIDAR studies.

Thanks to Gary Woffendin at Thermo Scientific, Hemel Hempstead, for analysing the human plasma and MIDAR preparations on the LTQ Orbitrap mass spectrometer and to Kathryn Lilley and Nick Bond from the University of Cambridge for their time and use of their instrument.

Most importantly, I would like to thank all of my family and friends for their continued love and support over the years. Especially my parents, for believing in me and my husband Ruben, for his patience and understanding.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AC	Alternating current
ACN	Acetonitrile
ACTH	Adrenocorticotrophic hormone
ALDOA	Fructose-bisphosphate aldolase A
AMT	Accurae mass and time tag
ATZ	Anilinothiozolinone
BMT	Basic mass tag
BSA	Bovine serum albumin
cDNA	Complementary DNA
CID	Collision induced dissociation
CK	Creatine kinase
CNBr	Cyanogen bromide
COFRADIC	Combined fractional diagonal chromatography
DC	Direct current
DDA	Data-dependent acquisition
DNA	Deoxyribonucleic acid
DPP IV	Dipeptidyl aminopeptidase IV
DTT	Dithiothreitol
EDTA	Ethylene diaminete tetra-acetate
ENOB	Beta enolase
ER	Endoplasmic reticulum
ESI	Electrospray ionisation
EST	Expressed sequence tag
FA	Formic acid
FT-ICR	Fourier transform ion cyclotron resonance
GP	Glycogen phosphorylase
HEPES	N-2-hydroxyethylpiperazine-N'-2-ethanesulfonic acid
HFM	Hollow-fibre-membrane
HP	Hydrophobicity
HPLC	High performance liquid chromatography
HSA	Human serum albumin
HUPO	Human proteome organisation
IAN	Iodoacetamide
ICAT	Isotope-coded affinity tag
ICR	Ion cyclotron resonance
IEF	Isoelectric focusing
Ig	Immunoglobulin
IMAC	Immobilised metal affinity chromatography
iTRAQ	Isobaric tags for relative and absolute quantification
LB	Luria broth
LC	Liquid chromatography
LDH	Lactate dehydrogenase
m/z	Mass to charge ratio
MALDI	Matrix-assisted laser desorption/ionisation

MDI	Methylenediphenyl 4,4'-diisocyanate
MetAP	Methionine aminopeptidase
MIDA	Mass isotopomer distribution analysis
MIDAR	Mass Isotomer Distribution Analysis of amino acid Residues
mRNA	Messenger RNA
MS	Mass spectrometry
MS/MS	Tandem MS
MudPIT	Multidimensional Protein Identification Technology
Nat	N-acetyltransferase
NHS	N-hydroxy succinimide
NME	N-terminal methionine excision
NMT	N-myristoyltransferase
NTPeps	N-terminal peptides
ORF	Open reading frame
PBS	Phosphate-buffered saline
PGAM	Phosphoglycerate mutase
PGM	Phosphoglucomutase-1
PITC	Isothiocyanate derivative phenylisothiocyanate
PK	Pyruvate kinase
PMF	Peptide mass fingerprinting
ppm	Parts per million
PPP	Plasma proteome project
PTC	Phenylthiocarbamyl
PTH	Phenylthiohydantoin
PTM	Post-translational modification
RNA	Ribonucleic acid
RF	Radio frequency
RP	Reverse phase
RT	Retention Time
SA	Serum albumin
SAX	Strong anion exchange
SCX	Strong cation exchange
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SP	Signal peptide
TCA	Trichloroacetic acid
TFA	Trifluoroacetic acid
TNBS	Trinitrobenzenesulfonic acid
ToF	Time of flight
TPI	Triose phosphate isomerase
Ub	Ubiquitin
v/v	Volume to volume ratio
w/v	Weight to volume ratio

ABSTRACT

Proteome complexity presents a major challenge in the field of proteomics. The majority of bottom-up methods begin with proteolysis, which increases the number of analytes in the mixture by about 30-50 fold. This level of complexity demands simplification, and there is an increasing requirement for strategies and reagents that reduce the complexity of a total proteome mixture.

It may be argued that when analysing a complete protein digest, for instance by standard shotgun methods, more peptides are analysed than strictly necessary. An efficient proteomic strategy simplifies the proteome while preserving most of the information necessary for comprehensive analysis. A practical approach to proteome simplification is to target a specific structural region of the protein molecule. The ultimate simplification strategy would be to select a single signature peptide from each protein in the proteome.

The primary goal of this study was to develop an advanced strategy for proteome simplification that reduces each protein to its N-terminal most proteolytic peptide, termed 'positional proteomics'. Under these circumstances, knowledge of the location of the peptide in the protein (at the N-terminal end) is a powerful adjunct to effective identification. In brief, a complex protein mixture is N-acetylated in its native state and proteolysed to generate a mixture of acetylated (blocked) N-terminal peptides and unmodified internal peptides. Subsequent incubation of the peptide mixture with an immobilised amine scavenger, results in binding of unblocked α-amino groups i.e. all internal peptides. The unbound N-terminal peptides are separated from the Sepharose and analysed with out further treatment.

N-terminal regions of protein molecules are, in general, poorly represented within published datasets due to a plethora of modification events which can occur both co- and post-translationally (e.g. N^{α}-acetylation, methionine excision, signal peptide excision and exopeptidase activity). The positional proteomics strategy provides method to characterise the true N-terminal region of proteins, identifying potential signal peptide cleavage sites, truncation by exopeptidase activity and modification by N^{α}-acetylation.

Once developed, this strategy was applied to the soluble fractions of three complex proteomes (*E. coli, S. cerevisiae* and mouse liver) in order to rapidly characterise and define the true N-termini of large numbers of proteins.

The issue of sample complexity is further complicated by the dynamic expression range of proteins in most biological mixtures. The profile of N-terminal peptides from human plasma allows identification of proteins that are previously obscured by peptides from other highly abundant proteins. N-terminal profiling of human plasma, in combination with protein normalisation, led to the detection of over fifty proteins, including low abundance proteins and potential diagnostic biomarkers.

The ability to determine the frequency of a specific amino acid within a peptide sequence will provide an additional parameter for database searching. In an approach termed MIDAR (mass isotopomer distribution analysis of amino acid residues), two labelled variants of acetic anhydride, separated by 1Da, were mixed asymmetrically (10% of the lighter, and 90% of the heavier variant). When peptides are labelled with this reagent, the isotope pattern of the products is complex, but the ratio of two specific ions gives a precise measure of the number of amino groups in each peptide. The approach is generally applicable to any peptide-based proteome analysis, but is used here as part of a positional proteomics based analysis. Knowledge of the number of amino groups can provide valuable information that can be used as an added parameter in accurate mass and retention time strategies.

1. INTRODUCTION TO PROTEOME SIMPLIFICATION

1.1 PROTEOMICS OVERVIEW

The rapid pace of genome sequencing efforts since 1995 has led to the complete, or almost complete, sequencing of over 1218 genomes, of which 792 are Bacteria, 136 Eukarya, 49 Archaea and 241 Viruses. The majority of these sequences are accessible in the public domain (http://ergo.integratedgenomics.com/ERGO/) (Overbeek *et al.,* 2003). The present challenge is to determine the function of the genes contained within these genomes and link their function to phenotype. The quest to elucidate gene function has led to the analysis of expression levels of the components that make up a biological system, most importantly, mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics). Recent advances in protein separation sciences, coupled with mass spectrometry, have laid the groundwork for a comprehensive analysis of proteins. Because proteins control the majority of biochemical processes, either through interaction with other proteins or between proteins and their substrates (Matthews *et al.,* 2001; Walhout and Vidal, 2001) it is anticipated that new strategies which rapidly and efficiently characterise proteins, will provide valuable insights into understanding gene function.

The term proteome was first used in 1994 and can be described as the global collection of proteins produced by an organism. Proteomics refers to the branch of discovery science focusing on proteins (Wilkins *et al.,* 1996). Understanding the proteome, the structure and function of each protein, along with the complexities of protein-protein interactions (protein complexes), is necessary in order to fundamentally understand and appreciate cellular systems.

1.2 PROTEOME COMPLEXITY

As the proteome is encoded by the genome, one might expect a one-to–one correspondence between genes and proteins. However, proteomes are typically of higher complexity than genomes. Single cell prokaryotes and eukaryotes typically contain a few thousand protein coding genes, for example, *Escherichia coli* contains around 4,200 genes (Blattner *et al.,* 1997) and *Saccharomyces cerevisiae* has 6000 (Goffeau *et al.,* 1996). Recent estimates predict that the human genome is comprised of around 20,000 to 25,000 unique genes (International Human Genome Sequencing Consortium, 2004), which is significantly lower

than previously thought (Ewing and Green, 2000). The increased physiological complexity seen in vertebrates is due to a complex set of processes used to regulate gene expression, and not simply an increased number of genes. Proteome complexity can be attributed to diversification at both the mRNA level and following translation of mRNA into protein, by covalent modification, in a process known as post-translational modification (PTM).

1.2.1 Diversity at the mRNA level

Diversification of mRNAs can arise by the use of alternative promoter sequences in the 5' upstream ends of mRNAs, leading to the generation of several homologous proteins of variable length, due to different starting points (also known as a nested set).

A much more significant factor in the generation of mRNA diversity is the alternative splicing of primary mRNA molecules. Alternative splicing is a process that increases transcriptome and proteome complexity by generating multiple mRNA products from a single gene. The mechanism involves processing of the mRNA transcript into different mRNA molecules by separating exons and reconnecting them in various ways to produce alternative ribonucleotide rearrangements (Sharp, 1985). Recent studies estimate that >60% of human and mouse genes undergo alternative splicing (Lander *et al.,* 2001; Sharov *et al.,* 2005; Johnson *et al.,* 2003). The process was originally thought to be the main mechanism for the generation of higher order diversity within eukaryotes (Ewing and Green, 2000). However, the use of expressed sequence tag (EST) analysis demonstrated that the amount of alternative splicing is comparable, with no large differences, between humans and other animals (Brett *et al.,* 2002). Therefore, it is thought that alternative splicing plays only a small role in the diversification of novel gene products. The difference in complexity seen throughout eukaryotic organisms is mainly a result of PTM of proteins.

1.2.2 Diversity at the protein level

The release of a completed polypeptide chain from a ribosome is often not the last physical step in the formation of a protein. In both prokaryotic and eukaryotic cells, a vast set of changes occur to amino acid residues in proteins after their emergence from the ribosome. These alterations (PTMs) serve to increase the diversity and heterogeneity of functional groups beyond that of the 20-22 types of amino acid incorporated into the nascent polypeptide chain (reviewed in Krishna and Wold, 1993).

Side chains in proteins, which potentially act as nucleophiles, are common targets for PTM; furthermore, the NH_2 and COO^- groups (at the N and C-terminus of protein molecules respectively), which can also potentially act as nucleophiles, provide common targets for modification.

Various covalent modifications often occur, either during or after assembly of the polypeptide chain. Most proteins undergo PTMs that play a pivotal role in many biological processes, such as gene expression, cell signalling and mitosis. PTMs can be divided into two classes: covalent addition of one or more groups, such as phosphoryl, acetyl or glycosyl, to one or more of the amino acid side chains and hydrolytic cleavage of one or more peptide bonds by the action of proteases. More than 200 different types of PTM are known and it is likely that additional ones are yet to be discovered (Creasy and Cottrell, 2004). All modified forms of each protein can potentially vary in abundance, activity or location within a cell. Modifications can be reversible or irreversible and can occur spontaneously or through enzymatic mechanisms.

The reversal of PTMs adds yet another level of complexity and functional regulation of proteomes. As a rule, reversible PTMs are most likely to be involved as on-off switches in the regulation of the biological activity of the proteins modified. In most cases, only the modified or unmodified form of the protein substrate retains biological activity, additionally, separate enzymes are required for the forward and reverse reactions. This type of modulation occurs through cascades of activation and inactivation which represent a powerful tool for living cells to amplify biological signals (Chock *et al.,* 1980). Reversible protein phosphorylation, principally on serine, threonine or tyrosine residues, is one of the most important and well-studied PTMs. In eukaryotes, protein phosphorylation can be a key regulatory event. Many enzymes and receptors are switched "on" or "off" by phosphorylation and dephosphorylation (Krebs and Beavo, 1979).

The majority of naturally occurring PTMs are non-reversible and lead to permanent changes in the protein sequence. Some of these permanent changes occur as a result of a spontaneous reaction without the involvement of enzymes. An example of a spontaneous non-enzymic process is deamidation of asparagine residues to a mixture of isoaspartate and aspartate (Rivers *et al.,* 2008). Deamidation of glutamine residues can occur but does so at a much lower rate and as a result, is limited to proteins with long half-lives (Robinson *et al.,* 1973). The role of deamidation is not clear, however, it has been proposed that deamidation may provide a signal for protein degradation, thereby regulating intracellular levels.

In addition to naturally occurring (*in vivo*) protein modifications, modifications can occur as a result of sample treatment and preparation (*in vitro*). Some of these modifications are relatively easy to recognise, such as methionine oxidation (Chao *et al.,* 1997) and cysteine carbamidomethylation (Wilm *et al.,* 1996), which are routinely incorporated into database search engines. However, the extent of *in vitro* protein modification is difficult to estimate as most are present substochiometrically (Wilmarth *et al.,* 2006).

1.2.3 Heterogeneity at the N-terminus

The work in this thesis is based predominantly on methods to characterise protein N-termini. Therefore, the focus of the following section will centre on complexity and heterogeneity at the N-terminal region. The α-amino group at the N-terminus of protein molecules is targeted by a variety of enzymatic processes which occur both during translation (co-translational) and following release of the polypeptide chain from the ribosome (post-translational). Three types of reactions prevail: hydrolytic cleavage to remove one or more amino acids, modification of the α-amino group and side chain specific changes.

Proteolytic events

N-terminal peptidases (aminopeptidases) function to systematically remove or 'trim' individual amino acid residues from the N-terminus of proteins, revealing new N-terminal residues which can potentially become substrates for enzymes that modify the α-amino group (Taylor, 1993).

All nascent peptide chains begin with a methionine residue, which is either unblocked (archaea and eukaryotes) or N-formylated (prokaryotes). In most cases, this residue is removed or excised by the action of methionine aminopeptidase (MetAP), which acts alone or in combination with peptide deformylase, which is responsible for N-formyl removal (Giglione *et al.,* 2000). N-terminal methionine excision (NME) is the major proteolytic pathway responsible for the diversity of amino acids in proteins and occurs in as much as 80% of proteins in any given proteome (Frottin *et al.,* 2006). NME is an irreversible co-translational mechanism that occurs before the nascent polypeptide chain is released from the ribosome (Espagne *et al.,* 2007). As a rule, the initial N-terminal methionine residue is removed from proteins in which alanine, cysteine, glycine, proline, serine threonine or valine is the penultimate residue (Prchal *et al.,* 1986). The lack of action of MetAP on proteins with large penultimate residues is a result of steric hindrance; In general, methionine is cleaved from penultimate residues having radii of gyration of 1.29 Å or less (Hirel *et al.,* 1989).

Functions of NME include an important role in protein turnover, in proteins targeted by the ubiquitin pathway (Giglione *et al.,* 2003). Furthermore, a recent study into N-terminal modification in eukaryotes suggests that NME acts as a marker, tagging the most abundant proteins in a biological system, which mostly have alanine in the penultimate position (Martinez *et al.,* 2008).

Some proteins containing proline as the antepenultimate (the third) residue are resistant to NME (Moerschell *et al.,* 1990). Methionine cleavage was completely inhibited from the Met-Val-Pro sequence of mutant human haemoglobin (Prchal *et al.,* 1986), which under normal circumstances would have undergone NME.

The signal hypothesis, originally proposed by Blobel and Sabatini in 1971 (Sabatini *et al.,* 1971), proposed that all mRNAs to be translated on bound polyribosomes contained a unique set of codons immediately preceding the start codon. When translated these codons result in a unique sequence of amino acid residues, the signal peptide (SP), at the N-terminal end of the nascent polypeptide. The SP sequence triggers the attachment of the ribosome to the endoplasmic reticulum (ER) membrane, providing appropriate conditions for translocation of the polypeptide chain across the membrane. SP removal is a co-translational process and occurs once the N-terminal portion of the nascent polypeptide has crossed the plasma membrane. This sequence of events is not restricted to secretory proteins (extracellular) but also applies to the synthesis of proteins synthesised in the cytoplasm that are destined for various cellular compartments (mitochondrion, plastid, endoplasmic reticulum, etc) and for membrane proteins. Proteins which remain in the cytoplasm do not possess signal peptides. SPs consist of a short transient sequence that is made up of three regions: a positively charged n-region, a hydrophobic h-region and a polar c-region leading up to the cleavage site (Figure 1.1; Schatz and Beckwith, 1990). The removal of signal peptides from secreted proteins is executed by the action of signal or transit peptidases located at the membrane of the targeted compartment (Jackson and Blobel, 1977). Mitochondrial targeting signals are rich in arginine, serine and alanine residues and lack aspartic acid and glutamic acid residues (von Heijne *et al.,* 1989). These peptides often form amphiphilic helices which are essential for their function (Roise *et al.,* 1988). Proteins that are targeted to the ER tend to have a high content of leucine residues (Nielsen *et al.,* 1996).

The prediction of signal peptides from primary sequence is a major component of automated protein annotation. Many software tools have been developed for cellular localisation prediction, using machine learning techniques such as neural networks and

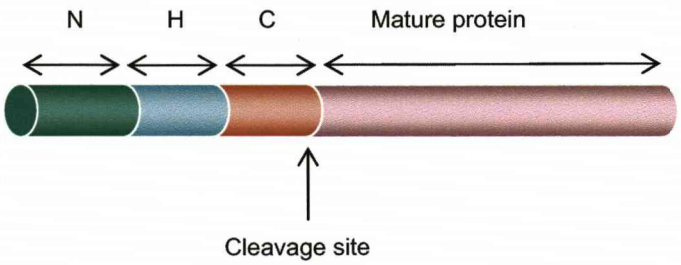


Figure 1.1. Signal peptide structure.
Signal peptides have a three-region design consisting of a positively charged N-terminal region (n-region), a hydrophobic central region (h-region) and a neutral, polar C-terminal region (c-region).

hidden Markov models (reviewed in Klee and Ellis, 2005). Most prediction programs predict intracellular signal peptides in proteins by identifying an N-terminal signal sequence, a signal sequence cleavage site, or a combination of both features, in a target sequence. Examples of signal peptide prediction algorithms include: SignalP (Bendtsen *et al.,* 2004), TargetP (Emanuelsson *et al.,* 2000) and PrediSi (Hiller *et al.,* 2004).

Propeptide cleavage is usually combined with SP cleavage, once the protein has reached its desired location. N-terminal propeptide removal occurs in combination with SP removal in the classic plasma protein serum albumin, which is cleaved at position 25, following the occurrence of two arginine residues (Dugaiczyk *et al.,* 1982). Propeptide cleavage often involves autocatalysis which can function to activate or switch on biological activity. Many enzymes are synthesised as inactivate precursors, otherwise known as zymogens (Stroud *et al.,* 1977). Trypsinogen is a zymogen which is converted to its active form (trypsin) by removal of a hexapeptide from the N-terminus. This cleavage reaction is mediated by the enzyme enteropeptidase, a membrane-bound serine protease. The enzyme hydrolyses a specific lysine-isoleucine peptide bond in the zymogen as it enters the duodenum from the pancreas (Maroux *et al.,* 1971). The small amount of trypsin produced in this process activates the conversion of more trypsinogen to trypsin, in a proteolytic reaction termed autoactivation. Conversion of trypsinogen to trypsin is of particular importance in the digestion process, as trypsin is responsible for the activation of all the pancreatic zymogens (pre-trypsinogen, chymotrypsinogen, proelastase and procarboxypeptidase; Bode *et al.,* 1978).

NME, SP cleavage and propeptide removal are all proteolytic mechanisms leading to the generation of new, mature N-termini. The main residues unmasked by NME are alanine and serine, which together account for 33% of the intracellular N-termini generated by endoproteolytic cleavage (Gevaert *et al.,* 2003).

N-blocking mechanisms

The α-amino group of a protein is also subject to various modification processes which result in the addition of a functional group. These modification processes include acetylation, myristoylation and propionylation.

Amino terminal modification by acetylation is extremely common in eukaryotes. Around 50% of the proteins in fungi are N^{α}-acetylated (Lee *et al.,* 1989a) and as many as 90% in animals (Polevoda and Sherman, 2000). N^{α}-acetylation is much less common in prokaryotes, although it does still occur (Charbaut *et al.,* 2002).

The process of N^{α}-acetylation, which occurs co-translationally, can modify the initiator methionine residue or the penultimate residue if the methionine is cleaved. The enzymes responsible for N^{α}-acetylaion in eukaryotes are N-acetyltransferases (Nats), which transfer acetyl groups from acetyl-CoA to a protein's α-amino group, via substitution with an active hydrogen atom. There are a total of three distinct Nat enzymes (NatA, NatB and NatC) each with different specificities. Unlike NME, the rules governing N^{α}-acetylation are not clear-cut. Eukaryotic proteins which are susceptible to N^{α}-acetylation have a variety of different N-terminal sequences with no simple consensus motif, and with no dependence on a single type of amino acid residue (Polevoda *et al.,* 1999). In accordance with NME, N^{α}-acetylation is partly dependent on the nature of the second encoded residue. Both alanine and serine are the most commonly N^{α}-acetylated amino acids, in addition to the uncleaved methionine residue when followed by either aspartic or glutamic acid (Flinta *et al.,* 1986). NatA typically modifies serine and alanine residues following methionine excision. In contrast, NatB and C have been shown to modify uncleaved methionine residues showing strong preference for proteins with a negatively charged (aspartic or glutamic acid) residue in the penultimate position.

The exact biological role of N^{α}-acetylation remains poorly understood. In *S. cerevisiae* the process is required for normal growth and mating (Lee *et al.,* 1989b), however in a separate study, the viability of *ard1*-Δ, *nat1*-Δ, *mak3*-Δ and *nat3*-Δ mutants suggests that the role of acetylation may be subtle and not absolute for most proteins (Polevoda and Sherman, 2003). It is likely that only a subset of proteins actually require this modification for activity or stability, whereas the remaining proteins are acetylated only because their N-terminal sequences correspond to the required consensus.

N-myristoylation, the covalent attachment of myristic acid (from myristoyl CoA) to the N-terminal glycine of proteins, is an irreversible co-translational process. Myristic acid, a 14 carbon saturated fatty acid (C14:0), is a rare fatty acid in biological systems, representing less that 1% of fatty acids in living cells (Boutin, 1997; Utsumi *et al.,* 2001). The process of attachment is an acylation reaction in which the acyl group is provided by a long chain fatty acid. The reaction is catalysed by the enzyme N-myristoyltransferase (NMT). The presence of myristoylate on the surface of a protein molecule increases lipophilicity allowing interaction with cellular membranes or other proteins. The majority of proteins targeted by NMT have an N-terminal consensus sequence of Met-Gly-X-X-X-Ser/Thre, although the initial methionine residue must be excised before attachment can occur (for a more detailed sequence motif, see Figure 1.2). N-myristoylation forms part of a multi-step process in which proteins are

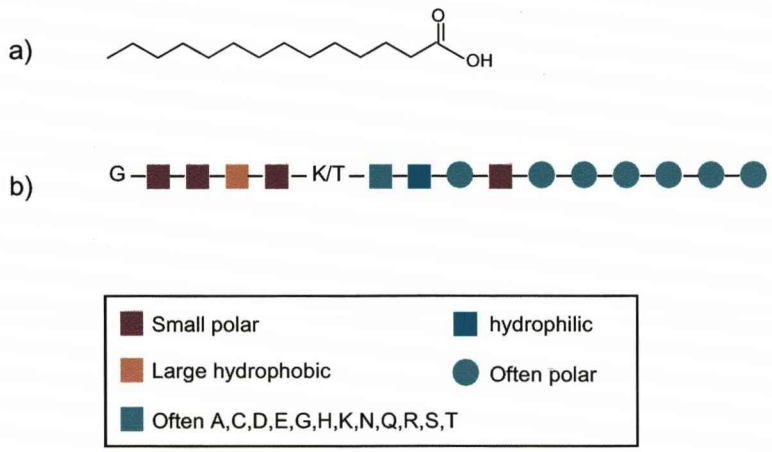


Figure 1.2. Myristoylation
a) Structure of myristic acid
b) Consensus sequence for recognition of protein substrates by NMT

transported to their functional locations. Many of the proteins modified by this process have crucial roles in disease progression. NMT is more active in colonic epithelial neoplasms than in corresponding normal-appearing colonic tissue (Magnuson *et al.,* 1995). Additionally, an increased expression of NMT was also observed in gall bladder carcinoma (Rajala *et al.,* 2000). Therefore, a greater understanding of the functions of NMT has important implications for cancer treatment and diagnosis.

Propionylation has recently been characterised as an N-terminal modification (Dormeyer *et al.,* 2007), and is thought to have a role in histone modification (Chen *et al.,* 2007).

N-terminal status and protein half-life

The N-terminal amino acid is thought to define the *in vivo* half life of a protein, according to the N-end rule (Varshavsky, 1992), a fundamental principle of regulated proteolysis that is conserved from bacteria to mammals. Although prokaryotes and eukaryotes employ a different set of components for degradation of N-end rule substrates, they share similar mechanisms of substrate recognition. In eukaryotes the N-end rule pathway forms part of the ubiquitin (Ub) system which is the pathway by which proteins are broken down into amino acids. Proteins are identified as being metabolically unstable and targeted for degradation, by the presence of degradation signals or degrons (Varshavsky, 1992). The essential component of a degron is a destabilising N-terminal residue, the N-degron. In eukaryotes the N-degron consists of not only a destabilising amino acid residue but also an accessible internal lysine residue which forms the site of the multi-ubiquitin chain (Varshavsky, 1997). Prior to degradation the Ub tag is attached to the ε-amino group of the internal lysine. Successive addition of activated Ub moieties to internal Lys48 of the previously attached molecule leads to generation of the polyubiquitin chain that is the degradation signal recognised by the 26S proteasome (Ciechanover and Schwartz, 1989). The ability of aminopeptidases to excise N-terminal residues could potentially influence a protein's interaction with the N-end rule machinery and modulate stability. If the aminopeptidase, for example MAP 1, revealed a destabilising or stabilising residue as the new N-terminus it could have a profound effect on the stability of the protein.

For most N-end rule substrates the initial Ub molecule is attached to the protein through its C-terminal Gly76 residue to the ε-amino group on the internal lysine. More recent findings indicate that, for some proteins, the first Ub molecule is attached to the α-amino group

on the N-terminus of the protein (Giglione *et al.,* 2003; Ciechanover and Ben-Saadon, 2004). In contrast to the traditional N-end rule pathway, N-terminal ubiquitination requires a free N-terminal amino acid to accept an Ub moiety.

1.3 Dynamic range of protein expression

The issue of proteome complexity is further complicated by the wide range of protein abundance present in most biological samples, referred to as the dynamic expression range (Corthals *et al.,* 2000). The dynamic range problem, in which the higher abundance proteins in the sample impede detection of the lower abundance proteins, presents a major concern for proteomic analysis. Proteins exhibit a very wide range in concentration, with a dynamic range of 10^5 in bacteria to 10^7–10^8 in human cells (Corthals *et al.,* 2000). Since there is no technique to amplify low-abundance proteins (comparable to the polymerase chain reaction for nucleic acids), both gel based and mass spectrometry methods often fail to detect the lower abundance complement (Kenyon *et al.,* 2002).

Substantial efforts have been made towards the global analysis of protein expression and localisation in the model organism *S. cerevisiae* which contains in the region of 6000 genes (Ghaemmaghami *et al.,* 2003; Huh *et al.,* 2003). The groups of O'shea and Weissman have utilised the availability of high-throughput technologies and the complete genome sequence of *S. cerevisiae* to systematically analyse a eukaryotic proteome. In this approach an alternative strategy for proteome characterisation was adopted. The groups created *S. cerevisiae* fusion libraries by tagging open reading frames (ORFs) with a high affinity epitope. These fusion proteins, which are expressed under the control of the organism's natural promoter, allow both immunodetection and immunopurification of the resulting fusion proteins. Creating this library allowed the researchers to demonstrate that around 80% of the yeast proteome is expressed at normal growth conditions, and also enabled the identification of miss-annotated genes. This study also showed that protein abundance can range from less than 50 to more than 10^6 molecules per cell. A total of 4,156 proteins (75% of the yeast proteome) were classified in this study, and subsequently organised into 22 distinct subcellular localisations. This global analysis of protein expression and localisation in a eukaryotic organism provides the most complete and physiologically relevant view to date. There are, however, concerns over the affect the epitope tag may have on protein turnover. When present on the N-terminal end of a protein the fusion tag may affect stability. If tagging affects the process of protein degradation then localisation of the protein could be compromised.

The dynamic range issue is of particular relevance when considering the proteome of blood, which is considered to be the most complex proteome in the human body (Anderson and Anderson, 2002). Virtually all cells in an organism come into contact with the blood, either directly or through tissue or biological fluids, as a consequence blood plasma contains not only characteristic plasma proteins but potentially every protein present in the body. Human serum or plasma, with a typical protein concentration of between 60 and 80 mg/ml is estimated to contain proteins spanning a concentration range of at least ten orders of magnitude (Anderson and Anderson, 2002). The six most abundant proteins (albumin, transferrin, haptoglobulin, α-1-antitrypsin, IgG, and IgA) account for approximately 95% of the proteome, with albumin alone representing 50% of the total protein content (Tirumalai *et al.,* 2003; Figure 1.3). The most clinically relevant proteins in plasma, i.e. biomarkers, are likely to occupy the low concentration range (present at ng/ml to pg/ml; Issaq *et al.,* 2007). Considering the nominal dynamic range of a mass spectrometer is at best three to four orders of magnitude, it is apparent that most proteomic studies will result in the identification of only the highly abundant (most likely the least interesting) portion of the plasma proteome. To have any chance of obtaining a comprehensive analysis of plasma proteins it is necessary to develop strategies to overcome the issue of sample complexity and dynamic range associated with such a complex biological sample.

The Plasma Proteome Project (PPP) was initiated by the Human Proteome Organisation (HUPO) in 2002 and involved a collaboration of 55 participating laboratories worldwide. In the space of three years substantial datasets were created and utilised to construct an integrated database of proteins identified in human plasma. The Core Dataset consisted of 3020 proteins identified by two or more peptide matches (Omenn *et al.,* 2005). This dataset has provided a strong foundation for future studies using human plasma and has developed understanding of both complexity and dynamic range.

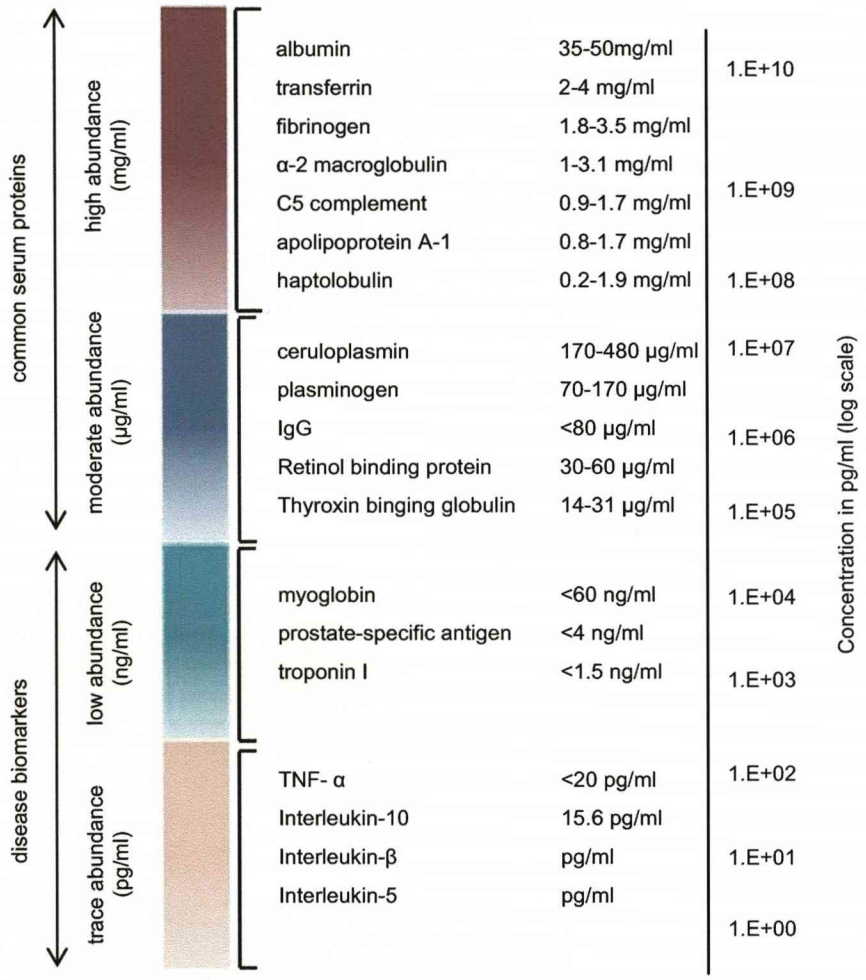


Figure 1.3. Protein concentration ranges in human plasma.
Serum and plasma are thought to contain a very complex set of protein signatures from most cell types in the body; however, these signature proteins are present at far lower concentrations than the classical plasma proteins.

1.4 Mass spectrometry

Mass spectrometry (MS) is one of the most informative techniques, widely used today for protein characterisation. MS based proteomics has been made possible by the availability of gene and genome sequence databases coupled with technical advances in the discovery and development of protein ionisation methods.

1.4.1 Ionisation methods

Ionisation techniques are a key requirement to convert molecules to gas-phase ions so that they can then be separated by electric or magnetic fields. For proteomics, the two most widely used ionisation techniques at present are electrospray ionisation (ESI; Whitehouse *et al.*, 1985) and matrix-assisted laser desorption/ionisation (MALDI; Karas and Hillenkamp, 1988). ESI ionises the molecules in-solution and is therefore readily coupled to liquid based (e.g. chromatographic) separation tools. In contrast, MALDI sublimates and ionises the samples out of dry, crystalline matrix via laser pulses (Figure 1.4). Both methods of ionisation are fundamentally the same, as the analyte mixtures are transferred into the gas phase and ionised. MALDI-MS is most commonly used to analyse relatively simple peptide mixtures, whereas integrated liquid chromatography ESI-MS systems (LC-ESI-MS) are preferred for the analysis of complex samples. Both ESI and MALDI are known as "soft" ionisation techniques which relates to the nature in which the ions are created. In these techniques peptide ions are generated with low internal energy and as a result undergo little fragmentation. In order to obtain structural information from an ion it is necessary to apply more aggressive ionisation methods such as electron ionisation which is described as a "hard" ionisation technique.

1.4.2 Mass analysers

Once the ions have been produced, they must be separated according to their mass to charge ratios (m/z). The most common types of mass analyser are time of flight (ToF), quadrupole, ion trap, ion cyclotron resonance (ICR) and more recently, Orbitrap mass analysers (Figure 1.5).

ToF analysers boost ions to the same kinetic energy by passage through an electric field and measure the times they take to reach the detector. While the nominal kinetic energy of all the ions is the same, the resultant velocity is different, thereby causing lighter ions (and also more highly charged ions) to reach the detector first (Figure 1.5a; Stephens, 1946).

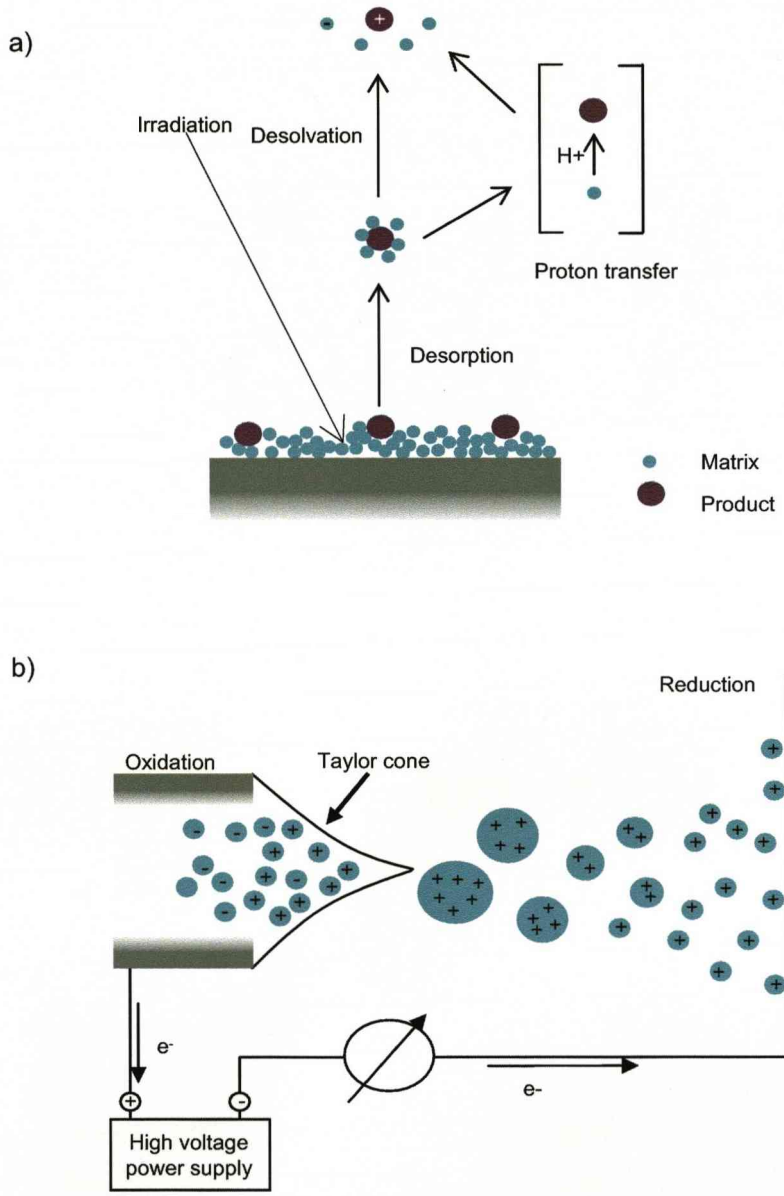
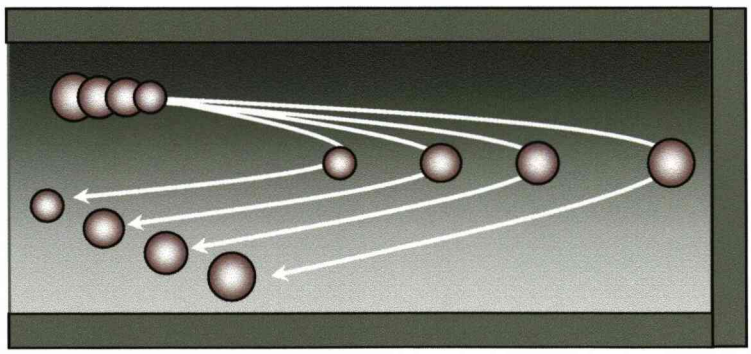


Figure 1.4. Ionisation techniques.
Matrix assisted laser desorption ionisation (MALDI; a) is used to generate gaseous phase analyte ions from the solid state. The analyte is spotted onto a stainless steel target and allowed to air dry. The analyte spot is then overlayed with matrix solution. The analyte and matrix co-crystallise to form a homogeneous layer, which is irradiated with a pulsed UV nitrogen laser. Energy adsorbed by the matrix is transferred to the analyte causing vaporisation of the analyte/matrix into a dense plume of molecules. A series of ion formation reactions occur, creating predominantly positively charged matrix and analyte ions. Electrospray ionisation (ESI; b) generates gaseous phase ions from the liquid phase. The sample solution is passed through a fine charged metallic capillary towards the entrance of the mass analyser. Because like charges repel, the liquid pushes itself out of the capillary and forms an aerosol. The aerosol is produced by the formation of a Taylor cone. As the solvent evaporates, the analyte molecules are forced closer together, repel each other and split into droplets. This process "Coulombic fission" repeats until the analyte is free of solvent and is a gaseous ion.

a)



b)

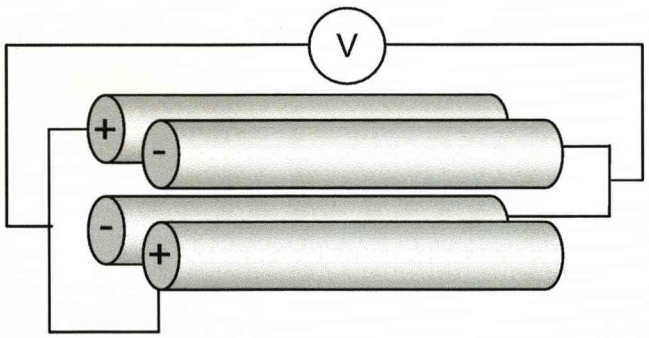
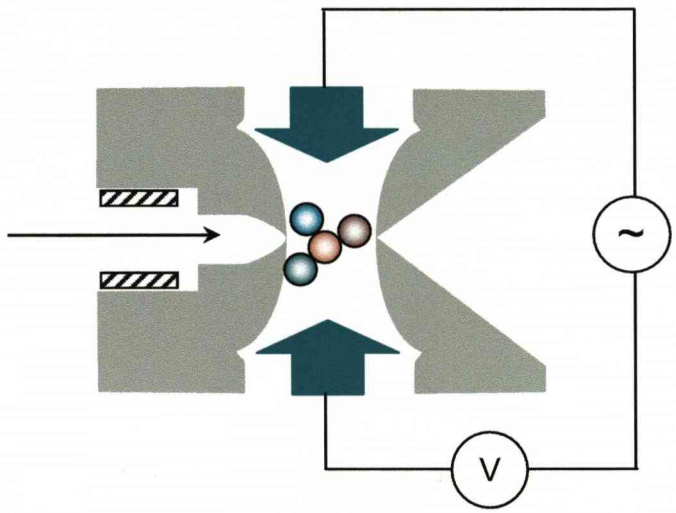
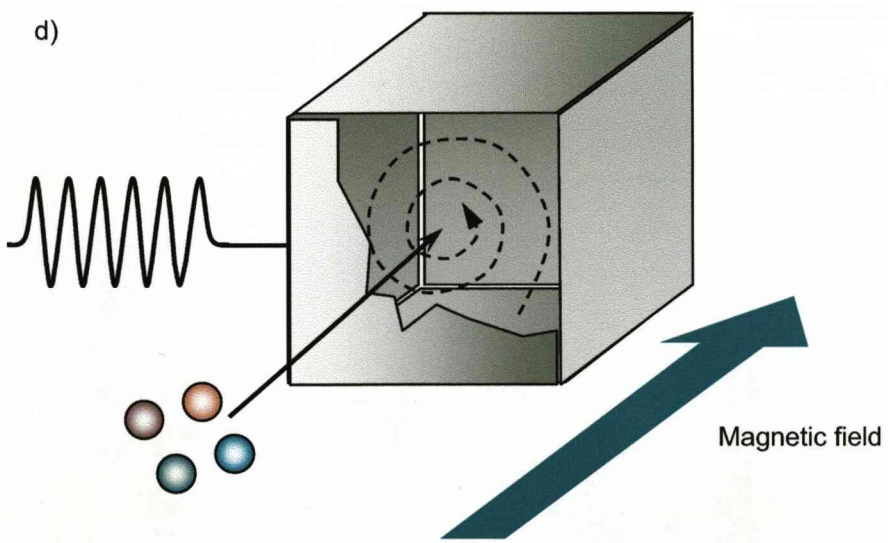


Figure 1.5. Mass Analysers.
The ToF mass analyser (a) measures the m/z of ions by pulsing them from the ionisation source into the flight tube. The m/z values are calculated by the time it takes the ions to hit the detector. In a quadrupole mass analyser (b), the correct magnitude of the radio frequency and direct current voltages applied to the rods allows ions of a single m/z to maintain stable trajectories from the ion source to the detector, whereas ions with different m/z values are unable to maintain stable trajectories. Ions in a quadrupole ion trap (c) maintain stable trajectories inside the device as a result of the application of a RF voltage to the ring electrode. Mass analysis is achieved by making ion trajectories unstable in a mass-selective manner. In an Fourier-transform ion-cyclotron resonance (FT-ICR; d), ions oscillate around the magnetic field at frequencies that are related to their m/z scales. As the ions oscillate near the top and bottom metal plates of the cubic trapping cell, they induce an alternating current that can be measured and then related to their m/z. Note that whereas the FT-ICR cell is small, it is in a high magnetic field (typically a super-conducting magnet), so the actual instrument size is large.

c)



d)



The quadrupole mass analyser is a "mass filter" consisting of four cylindrical rods arranged in parallel and a detector (Figure 1.5b). Combined direct current (DC) and radio frequency (RF) potentials on the quadrupole rods create an oscillating electric field which allows only ions with stable trajectories to pass through. By changing the applied electric field over time, it is possible to selectively transmit ions of a given m/z ratio to pass through and reach the detector (Horning *et al.*, 1977).

In the ion trap technique ions are focused using an electrostatic lensing system into the ion trap (Figure 1.5c). An electrostatic ion gate pulses open (-V) and closed (+V) to inject ions into the ion trap. The pulsing of the ion gate differentiates ion traps from "beam" instruments such as quadrupoles where ions continually enter the mass analyser. The time during which ions are allowed into the trap, termed the "ionisation period", is set to maximise signal while minimising space-charge effects, which are caused by too many ions in the trap causing distortion of the electrical fields leading to an overall reduction in performance (Kaiser *et al.*, 1991; Jonscher *et al.*, 1993).

ICR analysers (Baldeschwieler, 1968; Figure 1.5d) are based on the omegatron device (Hipple *et al.*, 1950) in which resonance is detected when ions starting from a fixed point achieve sufficiently large orbital radii to reach a fixed collector. The ICR analyser traps ions inside a cell consisting of four electrodes in a strong magnetic field (penning trap). Inside the trap, ions are excited, causing them to oscillate with a frequency perpendicular to the magnetic field, forming a "packet". The signal is detected as an image current on a pair of plates which the packet of ions passes close to as they cyclotron. Fourier Transform ICR (FTICR; Marshall *et al.*, 1998) MS is a high resolution technique in which masses are determined with very high accuracy.

Orbitrap mass analysers are the most recently introduced mass analysers (commercially available since 2005). Ions are electrostatically trapped in an orbit around a central, spindle-shaped electrode. The ion movement resembles a ring that oscillates along the axis of the spindle. This oscillation generates an image current in detector plates which is recorded. The frequencies of these image currents depend on the m/z ratios of the ions in the trap. Orbitraps share similar characteristics to FT-ICR instruments, in terms of resolution and mass accuracy but without the requirement of an expensive superconducting magnet (Scigelova and Makarov, 2006).

1.4.3 Peptide mass fingerprinting

One of the major applications of MALDI-ToF MS is peptide mass fingerprinting (PMF). This technique is used to identify single proteins by connecting a series of peptide masses to a known protein sequence within a database (Henzel *et al.,* 2003).The peptide mass is defined by the number and type of amino acids it is composed of, including any modifications the amino acids may have undergone. Information is obtained from a unique collection of peptides that occur when the molecule is digested with a proteolytic enzyme such as trypsin or Lys-C. Provided the digestion is complete the peptides produced will be of varying masses that are characteristic of that protein. The set of intact masses of the peptides produced by enzymatic digestion constitutes a unique set of masses or "fingerprint" for a specific protein. The experimental mass profile is matched against the theoretical masses obtained from the *in silico* digestion at the same enzyme cleavage sites of all protein amino acid sequences in the database. The matched proteins are then ranked according to the number of corresponding peptide masses within a given mass error tolerance. Proteins can be identified in this way with relatively high throughput (Pappin, 2003) and a high sensitivity, even below the femtomole range (Schuerenberg *et al.,* 2000).

1.4.4 Tandem MS

In order to generate data related to the amino acid sequence of a protein or peptide a variety of instruments have been developed to isolate ions, fragment them and determine the m/z of the fragments. These devices are collectively known as tandem mass spectrometers. These instruments differ in terms of overall design of mass resolution, mass accuracy, robustness and ease of operation.

Tandem MS (MS/MS) can be used to generate short amino acid sequences via fragmentation along the peptide backbone, by a technique known as collision induced dissociation (CID). The series of fragments produced differ in mass by a single amino acid and the differences can used to ascertain the amino acid sequence (Perkins *et al.,* 1999). The identification of peptides by MS/MS has substituted almost completely the more time consuming and relatively insensitive method of Edman degradation. Un-interrupted peptide CID fragmentation spectra are used to identify peptides and proteins with the help of programs like MASCOT (Perkins *et al.,* 1999) or SEQUEST (Sadygov *et al.,* 2004), which use mass fragmentation data to explore either protein databases (for example, SwissProt and NCBI) or

nucleotide data, such as the incomplete nucleotide sequences contained in EST databases (Choudhary *et al.*, 2001).

1.4.5 Improved MS capabilities for second generation proteomics

Together, PMF and MS/MS searches constitute the methodology used in most proteomics laboratories over the past decade. Many proteins have been identified in this way, for example characterisation of the *Haemophilus influenzae* proteome (Langen *et al.*, 2000) and isolation and characterisation of subnuclear compartments from *Trypanosoma brucei* (Rout and Field, 2001).

Proteomics studies are now moving away from "classical proteomic approaches" described above and shifting towards more sophisticated strategies in which high numbers of proteins are identified in a single experiment. High throughput protein identification, quantitative differences in protein expression and the global study of PTMs are among the applications desired by researchers. New methodologies are under development, which attempt to maximise the output from current instrumentation capabilities.

Traditionally the acquisition of MS/MS spectra is done in a data-dependent way through a process termed data-dependent acquisition (DDA; Belov *et al.*, 2001). A preliminary scan is carried out in MS mode and this scan is used to determine the ions chosen for fragmentation. Once an ion is selected (above a specified threshold) the instrument switches from full scan mode to MS/MS mode. The time required for the instrument to perform a full scan followed by one MS/MS scan is determined by the instruments duty cycle. In a typical DDA experiment, the most intense ions present at a given time are selected by the instrument for MS/MS. For complex samples this typically results in many lower abundance ions being ignored by the instrument, as a consequence these ions will remain uncharacterised and will not be represented in the dataset.

One of the most desired improvements in MS technology is shorter duty cycles, allowing more frequent sampling of ions. The acquisition of more MS/MS spectra per unit time would, in turn, improve dynamic range as lower abundant ions would be more likely to be sequenced. This improvement in scan speed coupled with other improvements in mass accuracy, sensitivity and pre-analysis separation methods will move researchers closer to achieving the goal of global protein characterisation. Currently, several different types of mass analysers (QToF, FTICR and Orbitrap) are competing with each other for high mass accuracy,

high throughput applications. This diversity means that the field of MS, although a century old, is still in the fast evolving phase.

1.5 PROTEIN SIMPLIFICATION

A complex protein mixture can be simplified through a series of separation or purification steps that exploit physicochemical properties between individual components within a mixture.

1.5.1 SDS-PAGE

Electrophoresis is a simplification technique that is routinely used to monitor the complexity of a protein mixture by separation on polyacrylamide gels. Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) is a technique used to separate proteins according to their electrophoretic mobility (Shapiro *et al.,* 1967). Prior to separation the protein molecules are denatured and coated with SDS so that they carry a large negative charge. During electrophoresis an electric field is applied through the gel causing the denatured proteins to migrate down the gel towards the positive electrode. The gel acts as a molecular sieve, separating the proteins on the basis of their molecular weight. Low molecular weight proteins migrate through the gel more freely than high molecular weight species and reach the bottom more rapidly. Separation of proteins on a polyacrylamide gel by mass alone is a one dimensional (1-D) separation. Two dimensional (2-D) electrophoresis begins with a 1-D step then targets a second characteristic of protein molecules for an additional phase of separation. In a typical 2-D electrophoresis experiment proteins are first resolved by their isoelectric point. To achieve this, proteins are first separated according to their charge (pI) by the process of isoelectric focusing (IEF; Laemmli, 1970). IEF pH gradients can be generated by adding ampholytes to an acrylamide gel. These are a mixture of amphoteric species with a range of pI values (O'Farrell, 1975). At pH values other than their isoelectric point, proteins will carry a charge. On application of an electric potential across the gel, the proteins migrate towards a point on the gel and accumulate at a position that corresponds to their isoelectric point. Once a protein reaches its isoelectric point its net charge will be 0 (i.e. neutral). There are issues regarding reproducibility between 2-D gels generated with carrier ampholyte IEF because of pH gradient instability. The problems of pH gradient instability and reproducibility were overcome by the introduction of immobilised pH gradients (IPG) for IEF (Bjellqvist *et al.,* 1982).

In the second dimension the proteins are separated by mass, as in 1-D SDS-PAGE. In this instance the electric field is applied at a 90° to the first field, which results in a gel with proteins spread out over the surface. The proteins can then be visualised using a variety of protein specific staining techniques including Coomassie blue and silver staining. Once visualised the individual protein spots can be excised for identification through various MS based approaches. 2-D gel profiles are particularly useful for comparative proteomic studies, in which proteins from control and experimental biological samples can be compared either by manual inspection or through a variety of computer based algorithms (Young *et al.*, 2004).

Although 2-D SDS-PAGE has been used for many years as a standard approach to protein separation, the drawbacks associated with this technique make it less suitable for high-throughput analysis of complete proteomes (Pieper, 2003; Hu *et al.*, 2007). Portions of proteomes including proteins with extremes in pI and molecular weight, low abundance proteins and membrane associated and bound proteins are rarely seen in a 2-D SDS-PAGE study. These factors lead to the visual under representation of the complexity of the protein sample on the gel. Another contributing factor to the lack of visual representation is poor dynamic range which severely restricts the ability to detect low-copy number and low abundance proteins. Also, the technique remains relatively low throughput due to the requirement for individual extraction, digestion and analysis of each spot from the gel (Washburn *et al.*, 2001).

1.5.2 High-performance liquid chromatography

High-performance liquid chromatography (HPLC; Regnier and Gooding, 1980) is a non-gel method of protein or peptide separation. The availability of different HPLC resins with different selectivity's offer alternative strategies for protein purification. The basic principle in HPLC is to flow the protein solution through a column packed with a specific matrix or resin. The material of choice will exploit a particular characteristic of the protein molecules causing individual proteins to react differently with the column. Proteins are separated by the time taken to flow through the column, or the conditions required to elute the protein from the column. Proteins are typically detected as they elute, by their absorbance at 280 and 214nm. There is a wide range of different chromatography methods routinely used in proteomics.

Reversed-phase

The most widely used mode of HPLC for separation of peptides or proteins is reversed phase liquid chromatography (RP-LC). The analyte interacts with the RP media based on the differences in magnitude of the hydrophobic interactions. The non-polar stationary phase is generally made up of hydrophobic alkyl chains ($-CH_2-CH_2-CH_2-CH_3$) that interact with the analyte. There are three common chain lengths, C4, C8, and C18. C4 is generally used for proteins and C18 and C8 are generally used to capture peptides or small molecules. With these stationary phases, retention time is longer for molecules which are more non-polar, while polar molecules elute more readily. Once bound, the analytes can be separated by running a linear gradient of an organic solvent such as acetonitrile (ACN) or methanol (MeOH).

Size exculsion

Size exclusion chromatography shares similarities with electrophoresis in that proteins are separated according to their size (Lathe and Ruthven, 1956; Porath and Flodin, 1959). The protein solution is passed through a column of porous beads. The large proteins in the mixture are not able to enter the beads and as a consequence pass straight through the column and elute first. The proteins that are small enough to access the beads will therefore take longer to elute from the column (Figure 1.6). The porosity of the beads can be adjusted to exclude all molecules above a given size. Sepharose and Sephadex are trade names of gel that are commonly available commercially in a broad range of porosities. A disadvantage of this method is its limited resolving power; however, it can be useful if the protein of interest is of extreme size relative to the other

Ion exchange

Individual proteins differ from one another in the proportions of the charged amino acids they contain. Lysine, arginine and histidine residues all carry positive charges, whereas glutamate and aspartate are negatively charged residues. The proportion of these charged amino acids will determine the net charge on a protein molecule. Protein separation by ion exchange chromatography relies on the interaction of charged molecules in the mobile phase (protein solution) with oppositely charged groups on the stationary phase (ion exchange matrix). The charges on these amino acids will vary depending on the pH of the solution they are dissolved in (Righetti and Caravaggio, 1976). Optimisation of the buffer conditions is therefore required in order to purify a specific protein for which the isoelectric point is known. Prior to chromatographic separation, a buffer is pumped through the column to equilibrate the

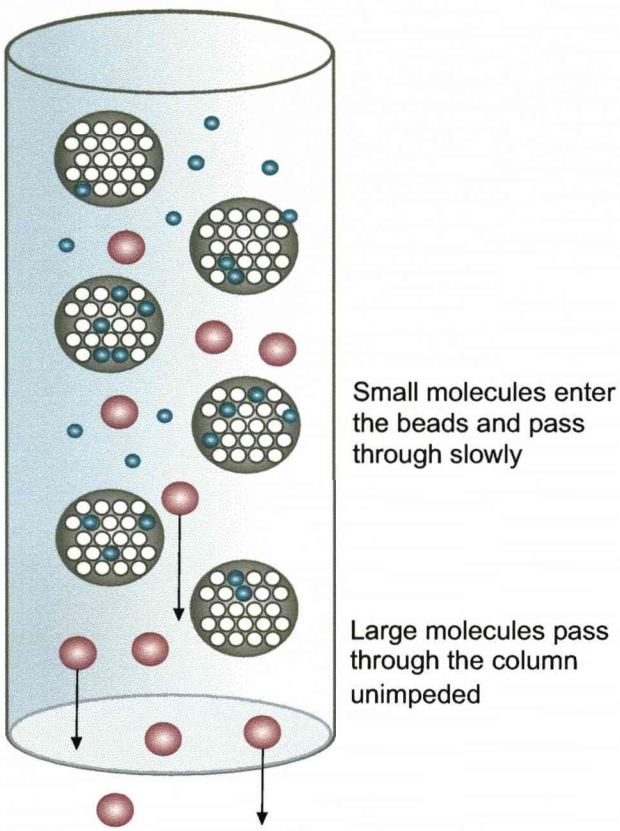


Figure 1.6. Size exclusion chromatography.
The column is filled with semi-solid beads of a polymeric gel that allow the admission of small molecules into their interior. When a mixture of proteins is introduced to the column, the small molecules are distributed through a larger volume of solvent than is available to the larger molecules. Consequently, the larger molecules move more rapidly through the column allowing fractionation of the mixture.

differentially charged ions on the ion exchange matrix. The protein mixture is introduced to the column in a solution of low salt. Under these conditions, charges on proteins bind to the charged matrix. To elute and separate the proteins, a gradient of increasing salt concentration is applied causing sequential dissociation of proteins from the column (Figure 1.7). The duration of binding or retention for each protein molecule will depend on the strength of interaction it shares with the matrix. Proteins carrying a high charge will interact strongly with the column causing them to remain associated for longer. Weakly charged molecules will elute from the column early on in the gradient followed by a succession of molecules of increasing charge.

There are two main types of resin used in ion exchange chromatography; cation exchange and anion exchange resins. Cation exchange resins have negatively charged groups on the surface and these are used to bind positively charged proteins. Anion exchange resins contain positively charged groups and these interact with negatively charged proteins. The type of resin used will depend on the pI of the protein or subset of proteins to be purified.

Affinity

The ability of specific protein molecules to interact with other structures forms the basis of affinity chromatography (Ostrove, 1990). This separation strategy uses specific ligands, attached to the chromatography resin, to target protein molecules for purification. The ligand is immobilised onto the chromatography resin in such a way that it retains its biological activity. Affinity columns can be used to select a specific protein, or class of proteins, from a complex protein mixture. Ligands typically function in a way similar to that of antibody-antigen interactions. This "lock and key" fit between the ligand and its target compound makes it highly specific, resulting in a single chromatographic peak, while the other components in the sample pass through the column, without interaction with the resin (Figure 1.8). An advantage of affinity chromatography over other techniques is its ability to purify target molecules present at low concentrations within a mixture. Two of the more widely used forms of affinity chromatography are Lectin and immunoaffinity based systems.

The main types of liquid chromatography methods are summarised in Table 1.1.

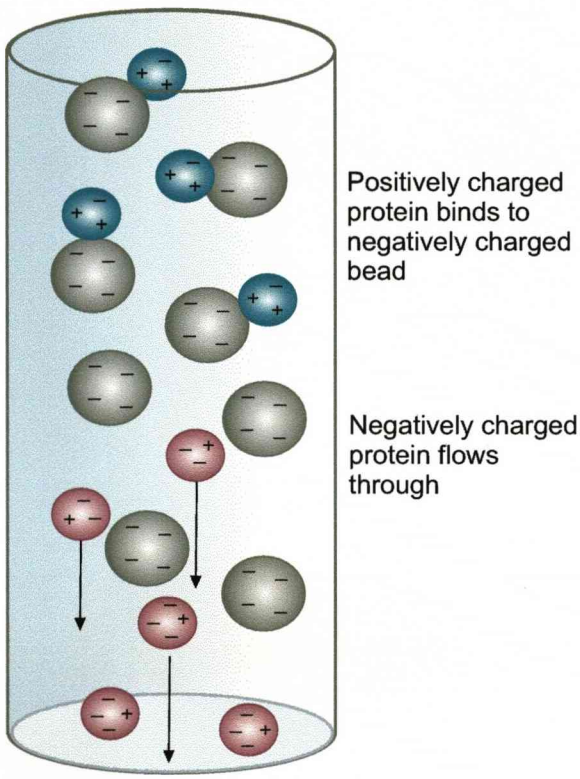


Figure 1.7. Ion exchange chromatography.
The column consists of beads displaying ionic functional groups. This type of chromatography is further subdivided into cation exchange and anion exchange (shown here) chromatography. When a mixture of proteins is applied to the column, molecules that possess a net charge opposite to that of the beads will bind, allowing their separation from proteins which have the same net charge as the beads, which will pass through. The bound proteins are eluted from the column in order of their binding affinity, by the application of a salt gradient.

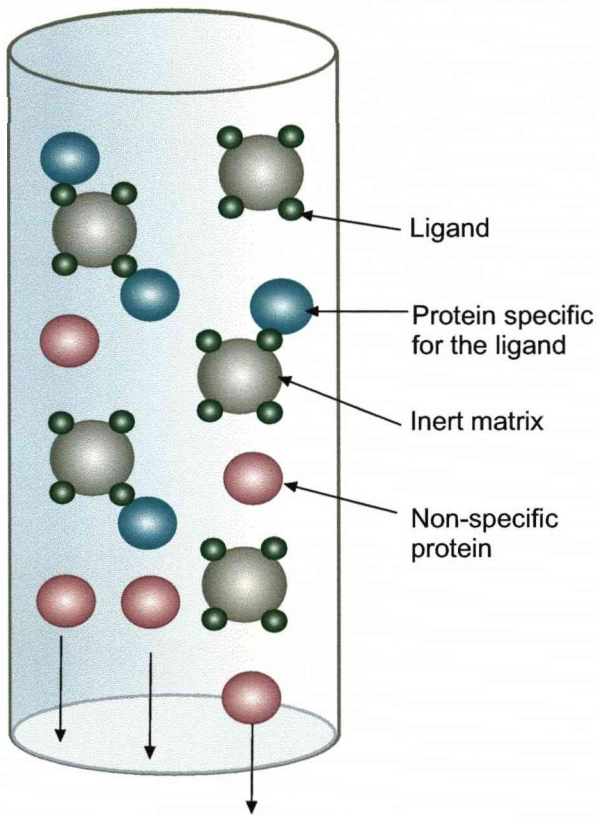


Figure 1.8. Affinity chromatography.
The column consists of an immunadsorbant matrix to which the ligand (green) has been coupled. The sample is passed over the matrix and the molecules that are specific to the ligand will bind with high affinity. Non-specific molecules will pass through the column unimpeded. The bound proteins can be eluted from the column using a variety of different reagents depending on the nature of the column (for example, salt gradient or urea).

Name	Abbreviation	Separation principle
Reversed-phase liquid chromatography	RPLC	Hydrophobicity
Size-exclusion chromatography	SEC	Size
Ion-exchange chromatography (strong cation exchange, strong anion exchange)	IEX (SCX, SAX)	Charge
Affinity chromatography	AC	Specific binding

Table 1.1. Basic types of liquid chromatography used in protein and peptide separation.

1.5.3 Depletion based simplification strategies

When complex biological samples contain large differences in protein abundance, separation by chromatography or 2-D SDS-PAGE will not be sufficient to characterise the low abundance proteins in the sample. To achieve a global proteomics analysis of a complex sample, it is therefore necessary to apply a combination of separation techniques to gain a comprehensive profile. The standard separation methods listed above cannot address problems of sample complexity and dynamic range without the incorporation of a more targeted approach. In samples that exhibit a wide dynamic range, an obvious strategy would be to remove or deplete the high abundance proteins and enrich the low abundance components prior to analysis. In the case of human plasma, in which half the total protein mass is accounted for by albumin, it would be beneficial to remove albumin from the sample, along with other abundant proteins such as immunoglobulin and transferrin. Current serum and plasma proteomic analysis strategies typically aim to overcome the dynamic range problem by employing a variety of protein-depletion methods. Once depletion has taken place the remaining proteins can be enriched and subjected to further proteomic analysis such as 2-D SDS-PAGE or gel-free chromatography based methods.

Non-specific methods

The simplest of these depletion techniques is non-specific depletion, in which chemicals are used to remove a major component. Many chemicals have been utilised for the removal of albumin from plasma. The dye Cibacron Blue has good binding affinity for albumin and is available at low cost (Hinerfeld *et al.,* 2004). A disadvantage of this method is the non-specific nature of the reagents leading to co-depletion of other components of the mixture, which could potentially include molecules of analytical importance.

Specific methods

Immunoaffinity depletion of albumin from plasma is far more efficient and specific than the chemical dye based methods. The use of affinity purified polyclonal antibodies to specifically remove albumin, Immunoglobulin, transferrin or fibrinogen was successful for the enrichment of low abundance proteins from human plasma samples (Tam *et al.,* 2004). There are currently at least seven commercially available affinity removal columns or kits for depletion of the six major proteins of human plasma. Although these kits effectively remove the high abundant components, the remaining proteins are still sufficiently abundant to prevent adequate detection of the low abundance proteins. Additionally, even though these methods

are greatly superior to the traditional chemical depletion methods, there still remains concern over co-depletion of low-abundance proteins. Albumin associates with a variety of other proteins that will be removed in the bound fraction and as a consequence will not be analysed. A combination of stepwise immunoglobulin G and albumin depletion followed by 2-D LC-MS/MS resulted in the identification of over 2000 proteins from a single plasma sample (Shen *et al.*, 2005). However, depletion of serum albumin also led to the co-depletion of a further 815 proteins. Depletion of IgG led to the co-depletion of a further 209 species. In a separate study, a commercially available albumin depletion kit was used to determine if removal of albumin would measurably reduce detection of lower abundance cytokine proteins in human plasma (Granger, 2005). The results from this study demonstrated that there may be a non-specific loss of cytokines following albumin depletion, which may also affect subsequent proteomic analysis.

1.5.4 Membrane removal

An alternative strategy to immunological depletion of abundant proteins in plasma is the membrane removal method. This strategy utilises membrane filtration to separate high-molecular weight proteins from low-molecular weight proteins in a sample. A recent strategy by Tanaka *et al* (Tanaka *et al.*, 2006) teamed this membrane removal method with a 3-D LC separation prior to MS/MS, resulting in the characterisation of around 1800 proteins. The initial separation stage used a device based on a hollow-fibre-membrane (HFM) system for the removal of high-molecular weight proteins. Many low-abundance proteins in plasma, including biomarkers, are smaller than albumin. Removal of the high-molecular weight fraction should therefore result in improved detection of low-abundance components. The HFM based device consists of a separation unit and a concentration unit, allowing for simultaneous separation and concentration of the low-molecular weight fraction. The device uses a multi-stage filtration process to ensure all high-molecular weight components are completely removed from the system. The resulting solution of low-molecular weight proteins was then subjected to a 3-D LC based separation which involved initial separation of the protein mixture using RP-LC followed by proteolysis then 2-D LC of the peptide mixture prior to MS/MS analysis.

1.5.5 Immobilised ligand libraries

A novel strategy for the normalisation of protein concentrations has recently been described by Righetti *et al* (Righetti and Boschetti, 2007; Righetti *et al.*, 2006; Sennels *et al.*, 2007). The approach involves the use of Protein Equalizer™ Technology to reduce protein concentration differences in complex protein samples. The methodology exploits the same physicochemical properties of proteins as those targeted in affinity chromatography. A complex library of ligands (immobilised onto the surface of beads) is used to effectively normalise the relative concentrations of proteins in a complex mixture. With sufficient diversity of ligands it should, in theory, be possible to capture every protein in the proteome. When mixed with the ligand library beads each unique ligand has the potential bind to a unique protein structure. Because bead capacity limits binding, high abundance proteins will rapidly saturate their available binding sites and most of the protein will remain unbound. In contrast, low abundance proteins will become concentrated on their specific ligands (Figure 1.9). It is this normalisation effect that serves to reduce the dynamic range of the sample and provides access to proteins that would be otherwise hidden. The bound proteins are retained at different strengths, allowing a stepwise elution process providing further simplification by sub-fractionation.

The ligands used are in the form of synthetically produced hexapeptides which are synthesised via a short spacer on poly(hydroxymethacrylate) beads. A batch containing millions of microscopic porous chromatographic beads are distributed into twenty separate vessels, corresponding to the number of building blocks (amino acids) used to create the library. Each bead vessel is assigned a different building block, which is chemically attached to the beads. The vessels are then mixed together, washed and split again into twenty individual batches and the building blocks attached. The whole process is then repeated until the desired sequence length is reached. In theory each bead, which contains millions of copies of the same ligand, will be different and the complete library will provide a ligand complementary to every protein present in a complex mixture.

This approach, in comparison to time consuming affinity depletion strategies has several advantages: Firstly, depletion techniques do not significantly enrich the trace components in the mixture, which often results in them being undetected and uncharacterised. Protein Equalizer™ technology normalises the concentration differences, which in turn serves to enrich trace components in relation to the high abundance species. Also, since ligand library beads do not deplete high abundance proteins, the problem of co-depletion is not an issue with this strategy. The use of this approach coupled to additional fractionation methods

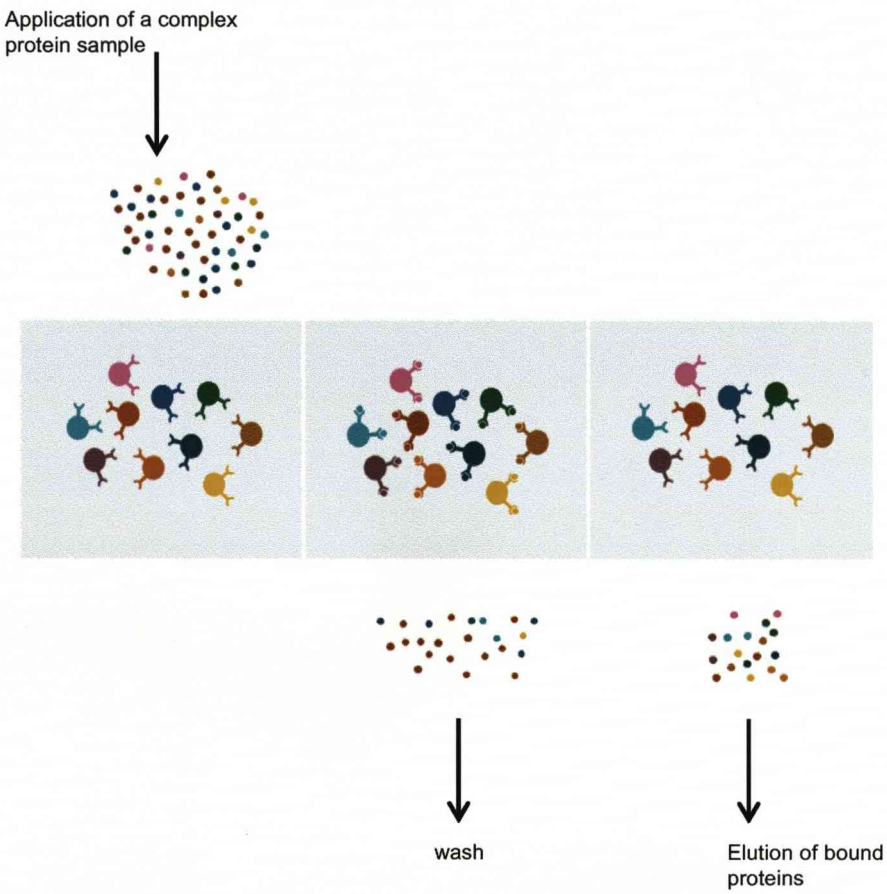


Figure 1.9. The mechanism of action of Protein Equalizer™ beads.
A complex protein sample is mixed with the ligand library causing the proteins in the sample to bind to their specific ligands. Abundant proteins quickly saturate their corresponding ligands and are removed from the mixture by subsequent wash steps. In contrast, low-abundance proteins are concentrated onto their binging partners. Following elution the resulting protein mixture exhibits reduced dynamic range of protein concentration.

resulted in an increased number of protein identifications and removes the problem of co-depletion associated with immunological based depletion techniques. However, due to the nature of the normalisation process, ligand library beads are not suitable for quantitative proteomic studies.

1.6 Peptide separation methods

Top-down proteomic approaches analyse intact protein molecules to identify both the protein and any potential structural modification. Due to the difficulty associated with processing intact mass data, these measurements are generally less effective for protein identification than peptide level analyses. The first step in top-down proteomics is fractionation at the protein level in order to reduce complexity and dynamic range (Righetti *et al.*, 2001). The main advantage of the top-down approach is that information regarding the intact protein structure is retained, which increases the likelihood of identifying protein isoform, polymorphism and PTMs.

The majority of in-solution based proteome studies are performed at the peptide level by bottom-up approaches (Link *et al.*, 1999; Washburn *et al.*, 2001). The disadvantage of bottom-up methods is the increased complexity that comes with digestion of proteins into peptides. It is feasible that a proteome consisting of 10,000 proteins will generate approximately 3.0×10^5 peptides, given that an average protein is digested into approximately 30 tryptic peptides. Additionally, proteolysis of a protein mixture leads to loss of connectivity between resulting peptides. Therefore, it appears satirical to introduce another level of complexity to the protein mixture by proteolysis. The rationale behind this strategy is the improved ability to analyse peptides over intact proteins. Peptides are characteristically more soluble than proteins; therefore, proteolysis will result in a wider range of available analytes.

Complexity demands simplification and in most "bottom-up" proteomics studies there is a point at which a transition must be made from 'protein space' into 'peptide space'. The transition is mediated by proteolysis, however, there is no set rule regarding when this step must occur (Figure 1.10). A key component in any proteomics analysis is the separation of peptides and proteins. Some proteomics studies aim to simplify a complex proteome, to the level of subproteomes (for example, a subcellular fraction) or individual proteins (such as proteins separated by 2-D gel electrophoresis) before proteolysis (Figure 1.10a). Other more global approaches begin with proteolysis of a complex sample, followed by peptide fractionation steps, for example HPLC, to achieve simplification (Figure 1.10b). In both cases,

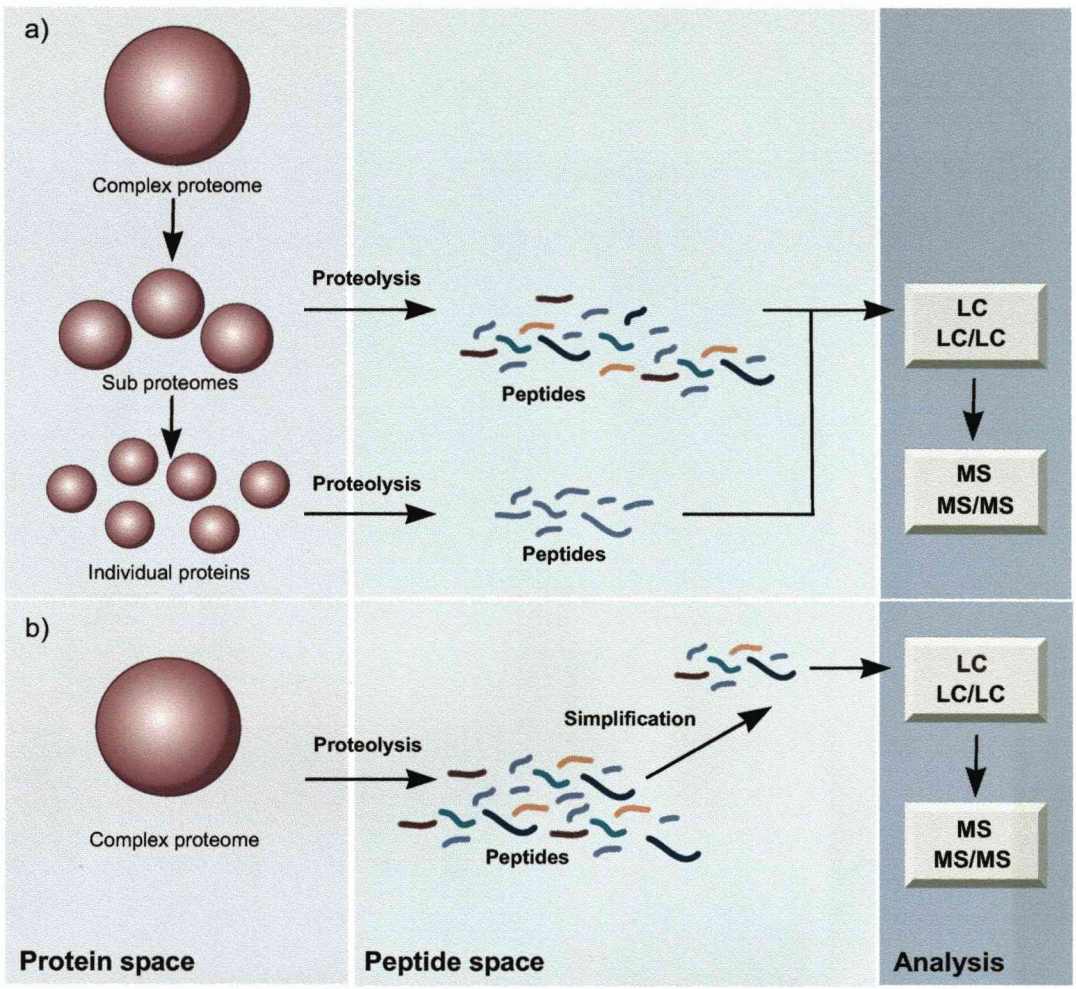


Figure 1.10. Outline of a standard approach to protein identification.
A complex proteome may be simplified using a variety of separation techniques in protein space (a). The resulting subproteomes can either be proteolysed directly or subjected to further separation into individual proteins. After proteolysis, the resulting peptides are analysed by mass spectrometry, with or without chromatographic separation, depending on the complexity of the peptide mixture. Proteolysis of a complete proteome creates a peptide mixture so complex that mass-spectrometric analysis is highly challenging. Simplification in peptide space (b) can be achieved be targeting specific structural regions on peptides, the mixture can be selectively purified in such a way that the majority of the proteome is discarded. The resulting simplified peptide mixture may be analysed by mass spectrometry, with or without chromatographic separation.

separation provides a way of reducing complexity and can also function as a method of delivering molecules to the mass spectrometer.

The term "shotgun proteomics" describes a bottom-up method in which HPLC is coupled to MS. A critical factor in the shotgun approach is the resolution of separation generated by the LC analysis. Straight forward LC-MS/MS analysis by shotgun proteomic approaches will only be capable of identifying a small portion of high abundance proteins in the sample. To dig deeper into the proteome it is necessary to apply more sophisticated simplification strategies. The application of multiple chromatographic stages prior to MS provides a way to address the issues of dynamic expression range and sample complexity. The combination of multidimensional separation methods with MS was first reported in 1997 (Opiteck *et al.,* 1997). A multidimensional separation includes two or more independent separation techniques (i.e. RP, IEX, SE and affinity chromatography), coupled for the analysis of an individual sample. The techniques used are complementary and achieve enhanced separation by exploiting different characteristics of peptides. To date, the most common multidimensional technique consists of two dimensions of chromatography (IEX and RP) allowing for enhanced resolution and peak capacity with the added advantage of solvent compatibility. By applying an additional dimension of peptide separation the quantity of peptides delivered to the mass spectrometer in a given time frame is reduced. The outcome or this is reduced ion suppression leading to an increased amount of peptides ionised at one time.

Later attempts at peptide separation by multidimensional chromatography led to the development of a more refined technique termed Multidimensional Protein Identification Technology or MudPIT, which can be described as a fully automated, online, coupled 2-D column MS/MS approach for the analysis of complex peptide mixtures (Link *et al.,* 1999; Washburn *et al.,* 2001). The peptide digest is loaded onto a column containing two or more chromatography resins in series. Peptides are eluted from each component of the column in a stepwise manner using salt solution plugs of increasing concentration. The first resin the peptides come into contact with is a SCX resin. Peptides bind the resin and are washed off sequentially using a salt step gradient. Eluted peptides then come into contact with the next component of the column which is typically RP resin. A second buffer with an increasing gradient of ACN is then used to wash peptides from the RP resin into the mass spectrometer (Figure 1.11).

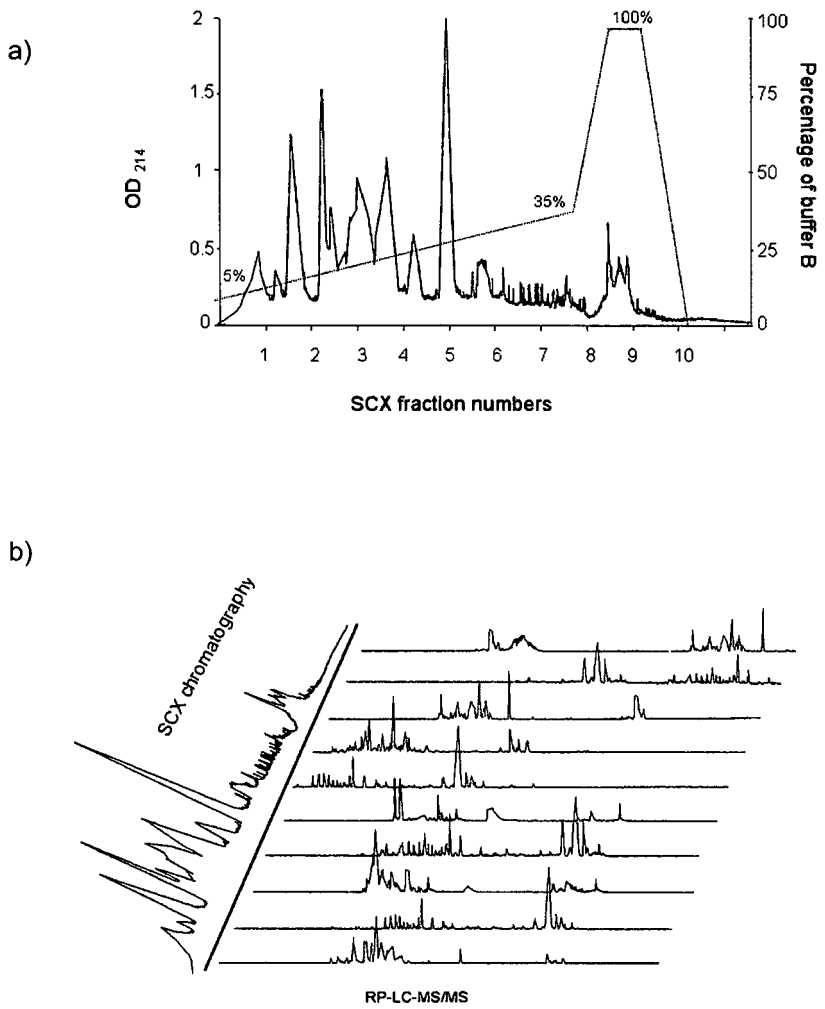


Figure 1.11. Multidimensional peptide chromatography.
Multi-dimensional peptide chromatography permits the analysis of thousands of proteins from a single sample. A highly complex peptide mixture is separated in the first dimension by SCX chromatography (a). The collected fractions are then analysed individually by RP LC-MS/MS (b) .

The MudPIT method of peptide separation has several benefits. Because an SDS-PAGE step is avoided, liquid phase protein samples can be easily proteolysed, loaded onto the 2-D chromatography column, and directly injected into the mass spectrometer. Additionally, when packing two types of media into a single column both types of media are exposed to all of the buffers used for equilibration, elution and washing steps. The advantage of using SCX coupled to RP media is that the two buffer systems are complementary. The salt solution plugs used to elute peptides from the SCX media are fully compatible with RP media. Furthermore, the addition of salt to the aqueous environment effectively increases the peptides binding affinity to the RP column by increasing the hydrophobicity of the solution (Mant and Hodges, 2002). Once the peptides have bound the RP media all salt containing buffers are sent to waste in the wash stage which ensures that no salt is allowed to enter the ion source of the mass spectrometer. Subsequent removal of peptides from the RP column by ACN mediates direct infusion of eluted peptides into mass spectrometer.

When applied to the analysis of the *S. cerevisiae* proteome, MudPIT resulted in the identification of 1,484 proteins, which included 131 proteins containing three or more predicted transmembrane domains (Washburn *et al.,* 2001). MudPIT has also been applied to the analysis of human plasma, where it was shown to increase identification of lower abundant proteins (Raida *et al.,* 1999).

Although multidimensional chromatographic methods are far superior to straightforward 1-D LC-MS/MS techniques, the number of peptides produced from a complex mixture of proteins generates extensive mass spectral data to be interpreted. Even with the benefit of 2-D chromatography, the mixture may deliver more peptides to the mass spectrometer than can be analysed in real time. Additionally, DDA is likely to direct MS/MS analysis to more abundant proteins, limiting dynamic range.

1.7 Proteolytic background

Shotgun proteomic approaches such as MudPIT are based on the assumption that each protein is present in a sample reproducibly and predictably generates a relatively small number of peptides. The peptides generated by an in-solution digest are restricted by the enzyme used and its cleavage specificity. The most commonly used enzyme for shotgun approaches is trypsin, which generates, on average 10 peptides for a stretch of 100 amino acids (based on the frequency of lysine and arginine residues). In shotgun based approaches, protein identification is typically achieved by searching multiple MS/MS files against protein

databases. When trypsin is chosen as a search parameter the peptides matched are restricted to full tryptic peptides. It is feasible that the products of a tryptic digest may contain many other components other than standard tryptic cleavages that are never seen, including partial trypic peptides and miscleavages. The implications for this hypothesis is a level of complexity that far exceeds current expectations for standard shotgun type approaches.

A recent study by Aebersold's group used a targeted peptide sequencing approach to investigate the true products of a tryptic digestion (Picotti *et al.*, 2007). The group used a set of well characterised proteins including β-lactoglobulin, carbonic anhydrase and β-casein which were individually proteolysed to perform an in-depth analysis. The targeted approach involved extensive characterisation of digestion products by MS/MS. Standard shotgun approaches employ a DDA strategy which targets the three most intense signals for fragmentation. This results in a large amount of unfragmented digestion products that are missed by the instrument. The group used an alternative sequencing approach to DDA which involved the construction of extensive inclusion lists which were subsequently used to trigger CID of the precursor ions. Samples were initially analysed using a high mass accuracy instrument in full scan mode. The data from this initial MS experiment was then extensively processed off-line in order to construct a comprehensive list of the species within the sample. The sample was then subjected to MS/MS using the inclusion list to direct fragmentation of all the ions detected in the sample. Each MS/MS sequence was subjected to manual assignment and only high quality matches were accepted.

An *in silico* digest of β-lactoglobulin produces 12 fully tryptic peptides containing more than five amino acids. In this study, using a conventional DDA sequencing approach 32 peptides were identified from the tryptic digest of β-lactoglobulin. In contrast, the targeted sequencing approach developed in this study resulted in the identification of 117 distinct peptides. Among these peptides were missed cleavages, amino acid substitutions and tryptic autolysis products, however, 89 β-lactoglobulin peptides were identified containing a non-tryptic terminus at the C-terminal residue.

The non-tryptic products generated in this experiment, although present in high numbers, were much less abundant than the fully tryptic peptides identified. For a complex biological sample, an extremely high number of low abundant peptides are expected to create a dense proteolytic background in the mass spectral data. For biological samples exhibiting large dynamic range, the proteolytic background from the high abundance proteins will serve to hide the signals of fully tryptic peptides from low abundance components in the sample.

This study revealed that the number of peptides observed from one protein is at least one order of magnitude higher than previously assumed. This unexpected complexity of tryptic digests has deep implications for the success of shotgun type proteomic approaches. Despite substantial efforts, the mapping of complex proteomes by shotgun methods is proving to be highly challenging. Findings from this study into proteolytic background may explain why shotgun approaches fail to provide in-depth coverage of complex proteomes such as plasma. This study highlights the requirement for alternative strategies to DDA and prompts the need for sample simplification techniques to simplify complex peptide mixtures.

1.8 Targeted analysis of peptides

It may be argued that when analysing a complete protein digest, for instance by standard shotgun methods, more peptides are analysed than strictly necessary. An efficient proteomic strategy simplifies the proteome while preserving most of the information necessary for comprehensive analysis. A variety of methods exist that exploit particular features of peptides in order to separate them from the total peptide pool. The isolation and analysis of a small but representative group of peptides has been shown to reduce sample complexity by 80% or more in a single separation step (Mirzaei and Regnier, 2005). Features such as the presence of specific amino acids (cysteine and histidine), PTMs and *in vitro* modifications have been targeted, along with the characteristics of peptides originating from protein termini. The identification procedure is greatly enhanced by knowledge derived from structure-based selection.

1.8.1 Targeting of specific amino acids

Low abundance amino acids such as cysteine and histidine are prime candidates for targeted peptide simplification. *In silico* analysis indicates that these amino acids occur in around 10-20% of tryptic peptides (Mirzaei and Regnier, 2005). The use of sequence-specific chemistry to capture peptides containing distinct amino acids provides a strategy to isolate a small but informative subset of peptides from, in most cases, each protein in a complex mixture.

Isolation of histidine containing peptides has been demonstrated using copper loaded immobilised metal affinity chromatography (Cu(II)IMAC; Raggiaschi *et al.*, 2005). In this study an entire tryptic digest was applied to the column and resulted in the purification of less than 5% of the total peptide mixture.

Cysteine containing peptides are typically selected by derivatisation of cysteine residues with a biotin reagent. Proteins are first reduced and alkylated with a biotinylated alkylating reagent followed by proteolysis. Peptides are then purified by a cation exchange chromatography step, which is necessary to remove excess reagents, the peptide mixture is then affinity purified using immobilised avidin to select cysteine containing peptides. Once desorbed from the column the purified peptides are characterised by LC-MS/MS based approaches.

Cysteine selection has also been used to study protein expression. A protocol for the specific enrichment and quantification of cysteine containing peptides developed by Aebersold and co-workers was first published in 1999 (Gygi *et al.*, 1999). The method is based on a group of novel chemical reagents termed isotope-coded affinity tags (ICATs) and MS/MS. The reagents consist of three components: a biotin tag, a linker containing either heavy or light stable isotope signatures and a thiol reactive group (Figure 12a). In the original ICAT reagent the heavy form contained eight deuteriums and the light form contained no deuteriums in the linker, this resulted in an 8 Da mass difference that could be differentiated by mass spectrometry. The next generation ICAT reagent, termed cleavable ICAT (cICAT), utilises carbon-13 instead of deuterium to create a heavy reagent. In this version of the ICAT reagent (Shiio and Aebersold, 2006), the mass difference between heavy and light isoforms is 9 Da which is due to the isotopically labeled linker that substitutes nine 13C atoms. cICAT also employs an acid-cleavable biotin group which facilitates removal of the bound peptides from the immobilised avidin.

In order to quantify cysteine containing peptides control and experimental samples are treated with the two isotopically distinct reagents and mixed prior to the proteolysis and ion exchange steps that precede avidin selection. If cleavable ICAT reagents are used, the biotin tag is released by acid cleavage after affinity chromatography. The isolated peptides are then separated and analysed by LC-MS/MS based methods. Each pair of light and heavy ICAT-labelled peptides are identical in terms of chemical composition and elution properties, but have a mass difference of 8 Da for original ICAT and 9 Da for cleavable ICAT (due to the incorporation carbon-13). Peaks corresponding to the same peptide are identified as doublets in the mass spectra due to the mass difference between the heavy and light reaction products. Changes in expression are quantified by analysing the LC MS data, the peak intensities correlate directly with the relative abundance of the proteins from the two samples (Figure 12b).

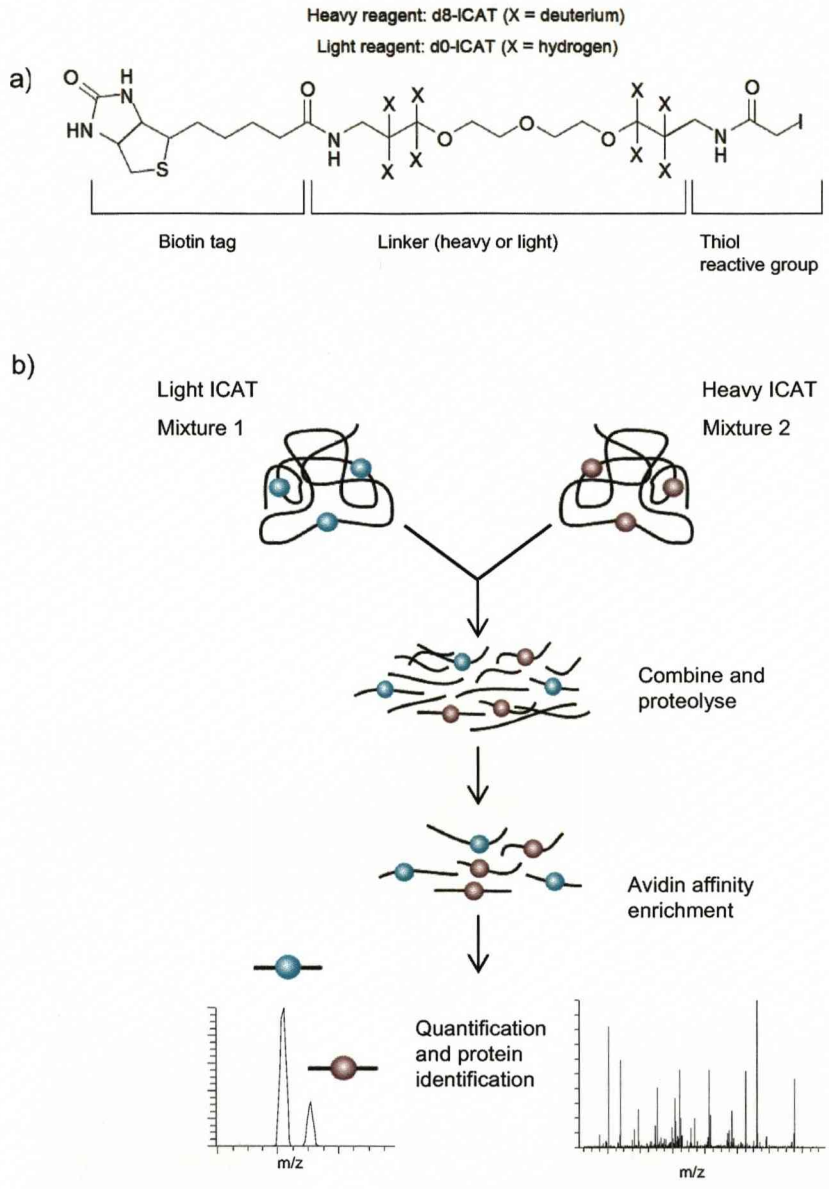


Figure 1.12. The ICAT strategy for isolation and quantification of cysteine containing peptides.
The ICAT reagent (a) consists of a thiol-reactive group which covalently binds to cysteine residues on proteins, a linker that can incorporate stable isotope signatures and a biotin tag used to purify the ICAT-labelled peptides by avidin affinity chromatography. There are two different tags, which are structurally identical except one has a linker containing eight hydrogen atoms, the other a linker with eight deuterium atoms. The two different samples are treated with different affinity tags (b). The samples are mixed and proteolysed to generated ICAT labeled peptides. The cysteine containing peptides are purified by avidin affinity chromatography, then characterised and quantified using MS based applications.

Present constraints to the ICAT technique centre on the issue of insufficient proteome coverage, which remains a concern for all MS based proteomics methods. A number of limitations have also been reported in the literature, including missed identification of proteins with few or no cysteine residues and lost information for PTMs. A small number of proteins naturally carry covalently bound biotin, in this case these proteins would also be selected by this process. The original ICAT reagent showed differential reversed-phase elution of identical peptides labelled with the hydrogen/deuterium isotope pairs, however, this issue was resolved when cICAT was developed (Wu *et al.,* 2006). Self alkylation of ICAT compounds can also occur which leads to instability of the ICAT reagents (Zhang *et al.,* 2005). In addition, fragmentation of the large affinity tag can further complicate mass spectra. Despite these issues ICAT technology has been applied to the analysis of whole-cell protein expression changes, and has provided unique insights into many biological systems.

Recently a new quantitative method, isobaric tags for relative and absolute quantification (iTRAQ) was developed. In contrast to ICAT, the introduction of stable isotopes using iTRAQ reagents occurs at the level of proteolytic peptides. The iTRAQ technology uses an NHS ester to modify primary amino groups on the peptides (Wiese, 2007). Although this method provides a powerful means of studying protein expression in multiple systems, the issue of simplification is not addressed. The iTRAQ quantification technique is discussed in detail in Section 3.3.

1.8.2 Targeting of post-translational modifications

Enrichment of phosphopeptides provides a way of simplifying a complex peptide mixture by discarding a large amount of peptides which are not phosphorylated and retaining those which are. Specific enrichment of phosphopeptides by immobilised metal affinity chromatography (IMAC) is a commonly used technique in proteomics. Using this approach it is possible to characterise hundreds of phosphopeptides from a whole-cell lysate in a single experiment (Ficarro *et al.,* 2002). IMAC exploits an immobilised transition state metal to bind phosphopeptides by virtue of their negative charge. IMAC is used routinely to purify phosphopeptides from peptide mixtures with iron and gallium being the metals of choice. Limitations to this method include the non-specific binding of carboxyl groups to the IMAC resins which compromises specificity of the system. Methyl esterification of carboxyl groups coupled with a low pH environment reduces the issue of non-specific binding.

Protein glycosylation is also acknowledged as a major PTM, with significant effects on protein folding, conformation distribution, stability and activity. These covalent modifications are known to play a key role cellular function and regulation (reviewed in Rademacher *et al.*, 1988), as well as being a key factor associated with disease by influencing regulatory and developmental processes. Carbohydrates in the form of aspargine-linked (N-linked) or serine/threonine (O-linked) oligosaccharides are major structural components of many cell surface and secreted proteins. To detect and characterise glycoslylated pepitdes with sufficient sensitivity it is often necessary to enrich for glycopeptides. As with other PTM's, digestion of glycoproteins often results in suppression of the glycosylated peptides by more abundant non-glycosylated peptides. A general approach for enrichment of glycoprotein and glycopeptides utilises the natural affinity that lectins have for glycans (Xiong *et al.*, 2003). Lectins are available that target either O- or N-linked oligosaccharides.

1.8.3 Diagonal chromatography

Diagonal chromatography is based on an old paper chromatography technique that has been updated for modern HPLC based proteomic applications. The original protocol was published in 1966 and was used in the characterisation of disulphide bridges in bovine chymotrypsinogen A (Brown and Hartley, 1966). The approach involved the initial chromatographic separation of a peptide mixture on a paper medium, which was then reacted with vapours of performic acid converting free cysteine to cysteic acid. Following this modification step the paper was rotated 90° and chromatographed in a second dimension using the same mobile phase. On separation, non-cysteinyl peptides migrated in the same way as in the first dimension, whereas cysteinyl peptides exhibited different mobility. The end result is that non-cysteinyl peptides appeared on a diagonal line and the modified cysteinyl residues migrated outside the line, hence the two types of peptides were resolved.

In recent years a variety of peptide separation methods have been developed that centre around diagonal RP liquid chromatography. These methods operate by targeting classes of peptides through specific chemical modification of the signature of interest. This technique has since been applied to the analysis of PTMs (Liu *et al.*, 2004). The method used was essentially 2-D HPLC with derivatisation. A set of tryptic peptides from a protein digest was separated by RP-HPLC, all fractions were collected, subjected to a derivatisation reaction and then re-chromatographed on the same column using the same conditions. This study also served to optimise the conditions for successful fractionation.

Combined fractional diagonal chromatography (COFRADIC) was first described in 2002 where it was used in the identification of more than 800 *E.coli* proteins (Gevaert *et al.*, 2002). This adaptation of the original diagonal chromatography method consists of three distinct stages: (1) a primary RP-HPLC fractionation of a complex peptide mixture; (2) a derivatisation or modification step on the peptide fractions; (3) a series of secondary RP-HPLC separations on the modified peptide fractions, using identical conditions to the primary fractionation (Figure 1.13). In the original application COFRADIC was used to isolate methionine containing peptides from a highly complex protein mixture. When examining model proteomes, methionine containing peptides provided the best representation of the predicted proteins. For the *E. coli* proteome, between 99.7 and 95.8% of the predicted proteins contained at least one methionine residue, in contrast only 85.4% of the proteins contained cysteine. This same trend in amino acid representation was observed in other model organisms.

In this investigation COFRADIC was used to select methionine peptides via oxidisation of methionine to Met-SO as the sorting criteria. The protocol began with a tryptic digest of a total, unfractionated cell lysate of *E. coli*. During the first chromatographic step, the peptides were separated and collected at appropriate time intervals. The individual fractions were then treated with hydrogen peroxide, leading to modification of the subset of methionine containing peptides. The specific conversion of methionine to its sulfoxide derivative occurs without affecting other susceptible amino acids like cysteine or tryptophan. Each fraction was then run under the same chromatographic conditions as the primary run, this separation is referred to as the secondary run. Since methionine sulfoxide is more hydrophobic than methionine, the modified peptides undergo a hydrophobic shift which is predictable if peptides contain one methionine. During the second COFRADIC dimension the oxidised methionyl peptides migrate out of the primary collection interval and are automatically collected for further MS/MS analysis.

COFRADIC has proven to be an extremely powerful technique for gel-free proteomics. One of the major advantages of this method over other targeted simplification strategies is its versatility. Diagonal chromatography is not restricted towards the analysis of one class of peptides, in contrast to techniques such as ICAT in which one given set of peptides is selected. By changing the key modification reaction, it is possible to tailor the process so that different classes of peptides are sorted and analysed. Applications for separation of proteins through virtue of cysteinyl (Gevaert, 2004), amino terminal (Gevaert, *et*

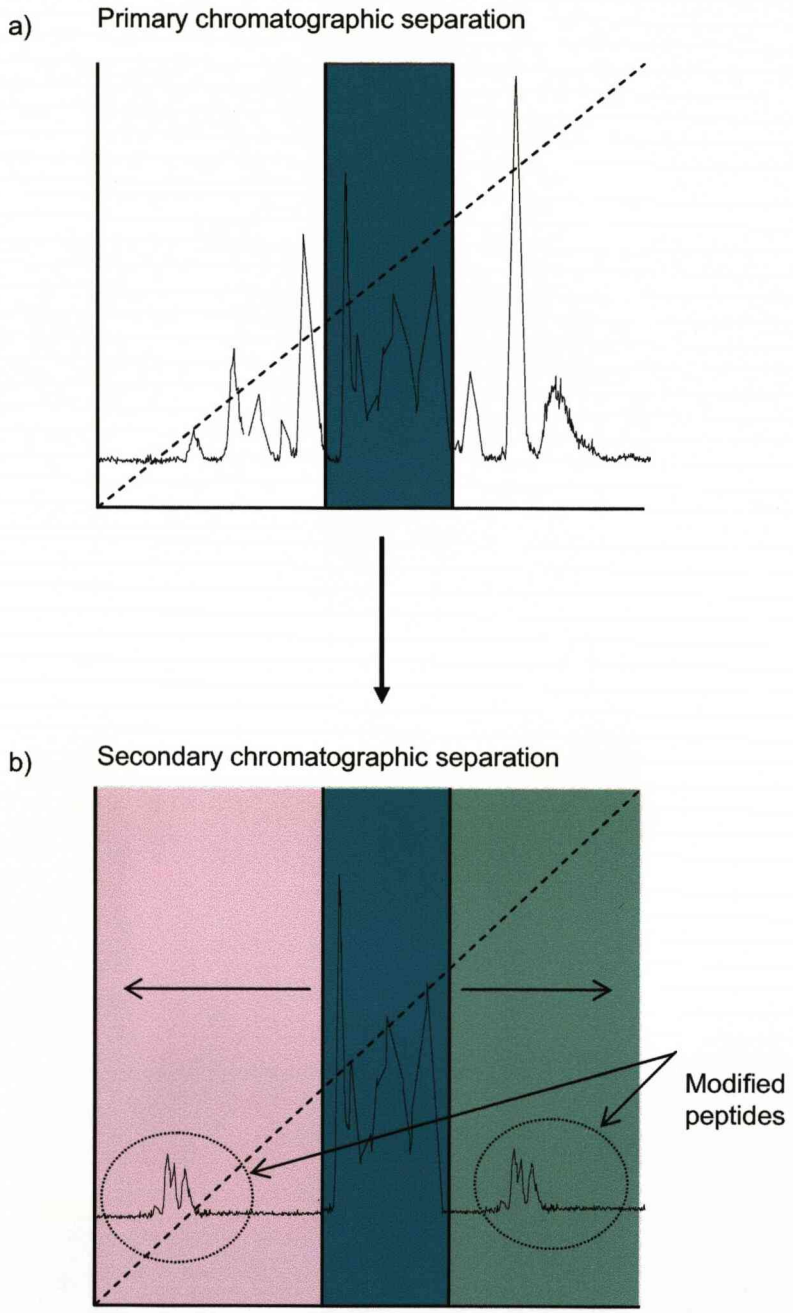


Figure 1.13. Peptide elution profiles during the primary and secondary run of COFRADIC.
During the primary run (a) several fractions are collected and treated with a specific modification reagent. During the secondary run (b) the modified peptides shift to earlier or later elution times than previously observed in the primary run. Unaltered peptides elute at the same interval.

al. 2003) and phosphorylated peptides (Gevaert *et al.*, 2006) have been reported in the literature.

1.9 Positional specific peptide isolation

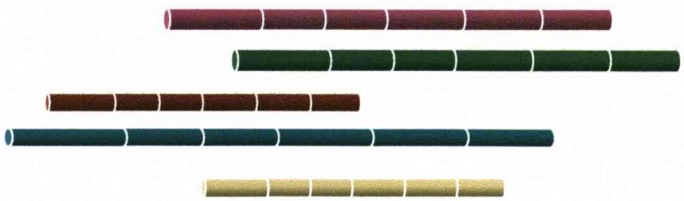
All of the peptide simplification techniques reviewed so far have the potential to extract more than one peptide for each protein and in some instances, no peptides will be recovered. Furthermore, because the peptide can have been derived from any part of the parent protein sequence, no information on the location of the peptide is obtained, which complicates the search strategy for identification proteomics.

The ultimate simplification strategy would be to select a single signature peptide from each protein in the proteome. Methods that isolate either the N-terminal or the C-terminal most peptide from a complex peptide mixture are the obvious choices for such a strategy. By isolating a specific peptide from a known location in a complex peptide mixture it is possible to achieve not only simplification but added information regarding the location of the peptide within the parent protein molecule (Figure 1.14).

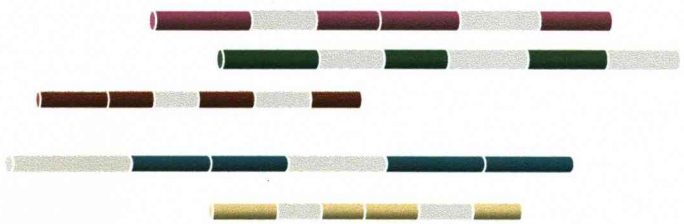
1.9.1 C-terminal peptide isolation strategies

Methods for recovery of C-terminal peptides are predominantly based on capture of internal tryptic peptides using immobilised anhydrotrypsin (Kasai, 1992; Sechi and Chait, 2000). A characteristic of tryptic digests is that all internal and N-terminal peptides terminate at basic residues (lysine or arginine). This method exploits this characteristic and uses the basicity of internal peptides as sorting criteria. Anhydrotrypsin is a catalytically inert form of trypsin containing a dehydroalanine residue in place of Ser-195 at the catalytic site. Although it has no catalytic activity, anhydrotrypsin retains strong affinity for peptides with C-terminal lysine and arginine residues. Anhydrotrypsin is synthesised by reacting trypsin with phenylmethylsulfonyl fluoride followed by potassium hydroxide treatment (Ako *et al.*, 1974). Incubation of a mixture of tryptic, Lys-C or Arg-C peptides with the immobilised anhydrotrypsin results in a covalent interaction with peptides terminating in a basic residue (i.e. N-terminal and internal peptides) leaving behind the C-terminal fragment. The anyhrotrypsin C-terminal enrichment protocol begins with proteolysis of a single protein or protein mixture with either trypsin or endopeptidase Lys-C (a combination of the two proteases may be used for enhanced coverage). The resulting peptide mixture is incubated with immobilised

(a) Target proteins



(b) "Shotgun" peptide analysis



(c) Positional proteomics

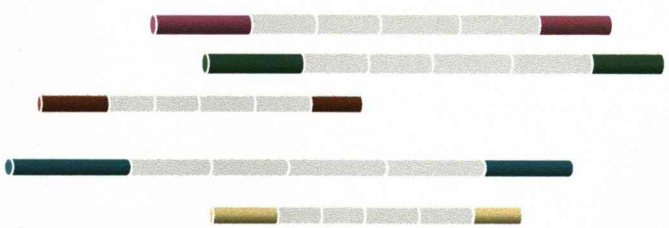


Figure 1.14. Strategies for targeted peptide simplification.
Proteolysis of a complex protein mixture generates large sets of peptides from each protein (a). Shotgun peptide analysis (b) randomly sequences a few peptides from each protein. Positional proteomic strategies achieve simplification by isolating specific peptides (N or C-terminal, depending on chemistry used) from each protein in the mixture.

anhydrotrypsin. The unbound fraction contains the purified C-terminal peptides that may be characterised by MS (Figure 1.15).

One problem potentially associated with this method is the failure to purify C-terminal peptides terminating with lysine or arginine. This issue can be addressed by incubation of the protein mixture (prior to proteolysis) with carboxypeptidase B, an exopeptidase known to catalyse the removal of basic amino acid residues from the C-terminus of proteins (Perryman *et al.*, 1984).

Despite a variety of potentially useful chemical modifications of the C-terminal amino acid including conversions of the carboxyl group to thiohydantoins, carboxylic acid esters, alcohols, acylureas, isothioureas, azides and hydrazides, researchers have found the published procedures challenging. The most likely avenue for C-terminal isolation by a chemical tagging strategy is through the utilisation of oxazolone chemistry (Nakazawa *et al.*, 2004; Yamaguchi *et al.*, 2006). Although modification of C-terminal groups on proteins is less effective than the established techniques for N-terminal modification (for example, Edman chemisry), it provides the sole means to deal with C-terminal carboxyl groups, discriminating against the same group on side chains. A major breakthrough in C-terminal peptide isolation could therefore be achieved through the use of oxazolone chemistry.

1.9.2 N-terminal peptide isolation strategies

There are many N-terminal isolation strategies available in the literature. These methods are typically based on a series of chemical modifications, proteolysis and a final sorting stage in which the N-terminal most peptide from each protein is selected. These methods vary in the types of chemistries involved and the way in which the targeted peptides are isolated from the mixture.

N-terminal COFRADIC

As previously described, COFRADIC is a highly versatile peptide fractionation technique that can be applied to a variety of classes of peptides. When isolating methionyl or cysteinyl peptides, the complexity of the analyte mixture is reduced to around five fold. In most global proteomics studies this will continue to generate too many peptides to analyse. By altering the sorting chemistry to specifically target N-terminal peptides it is possible, in theory, to limit each protein in the mixture to one characteristic signature peptide. COFRADIC has been adapted for the selection of N-terminal peptides from a complex peptide mixture

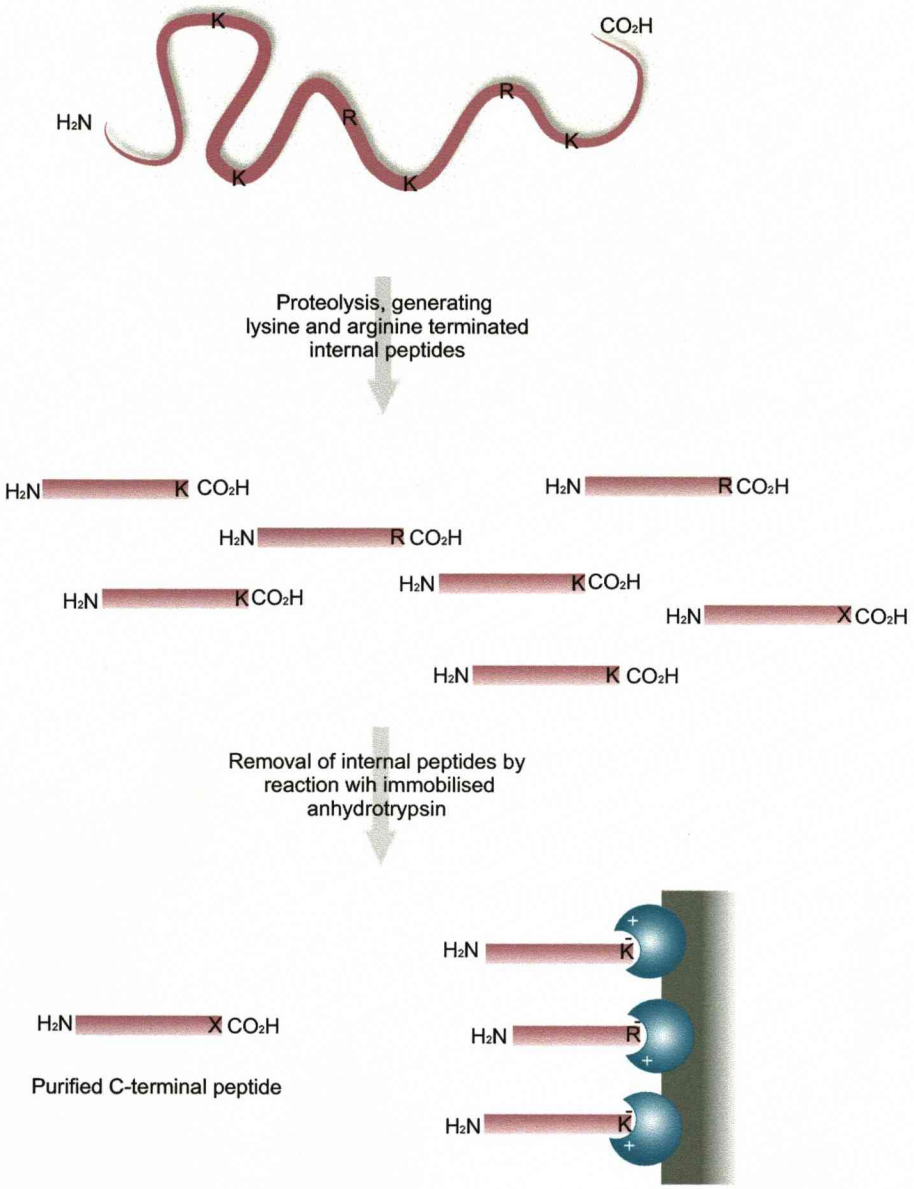


Figure 1.15. Purification of C-terminal peptides using immobilised anhydrotrypsin. The protein (or protein mixture) is proteolysed with trypsin generating a set of lysine and arginine terminated tryptic peptides. Internal peptides are removed from the mixture through incubation with immobilised anhydrotrypsin which binds to and retains basic (lysine and arginie terminated) peptides. The C-terminal peptide, which potentialy contains any amino acid residue (X) at the C-terminus, remains in the supernatant fraction (providing X is not arginine or lysine) and is retained for further analysis.

(Gevaert *et al.*, 2003). The rationale behind the sorting of N-terminal peptides is as follows: (1) reduction, alkylation and acetylation of a complex protein mixture; (2) digestion of the modified proteins; (3) RP-HPLC fractionation of the modified peptides (primary separation); (4) derivatisation of free α-amino groups on the newly formed peptides in the HPLC fractions using 2,4,6 - trinitrobenzenesulfonic acid (TNBS); (5) a second RP-HPLC separation of the modified peptide fractions using identical conditions to the primary run (secondary run).

Beginning with a whole-cell lysate, the cysteine residues are first reduced and alkylated. All exposed amino groups (α and ε) are then acetylated with N-hydroxysuccinimidyl (NHS) acetate and the modified protein mixture proteolysed with trypsin. Blocking of lysine residues by acetylation makes them inaccessible to trypsin, which will subsequently only cleave only at arginine residues. The resulting arginine terminated peptide mixture will consist of blocked N-terminal peptides mixed with internal peptides containing free amino groups on the newly generated N-termini. The complex mixture of acetylated peptides is then fractionated by RP HPLC, typically in 12 fractions (primary run). The peptide fractions are dried and reconstituted in an appropriate, non-amine containing, buffer. Each fraction contains a mixture of internal (unblocked) and N-terminal (blocked) peptides. Each fraction is then reacted with TNBS in which a trinitrophenyl group is placed upon the N-terminus of the internal peptides. Each TNBS treated primary fraction is then fractionated separately on the same column under identical separation conditions as the primary run. During a series of secondary COFRADIC separations, the TNBS-modified internal peptides exhibit increased hydrophobicity and as a consequence shift out of their primary collection interval. By excluding the acetylation stage in the protocol it is possible to isolate non-lysine containing N-terminal peptides from *in vivo* acetylated proteins. The entire COFRADIC N-terminal strategy is summarised in Figure 1.16.

This procedure was applied to the proteomic analysis of a cytosolic and membrane skeleton fraction of human thrombocytes. A total of 264 proteins and 78 *in vivo* acetylated proteins were identified. A problem associated with N-terminal peptide analysis by MS based approaches is the variable processing of protein N-termini. Therefore, it is not always possibly to identify processed N-terminal peptides using existing databases due to the incomplete knowledge of N-terminal boundaries. To successfully identify the whole range of N-terminal peptides it was necessary to generate a database containing sequentially trimmed Arg-C peptides. The use of this database in combination with established database search tools led to an increase in identification of around 50%.

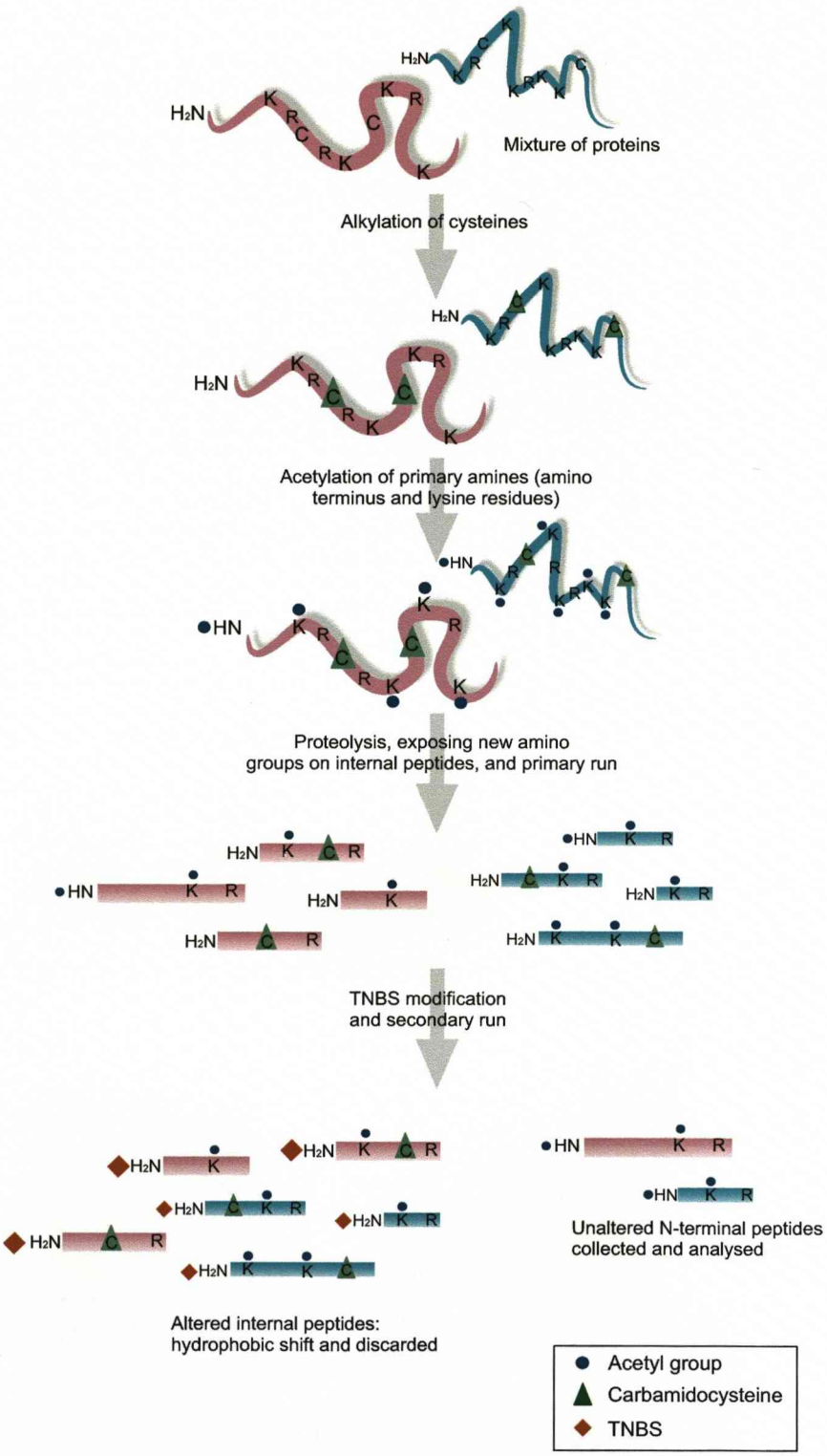


Figure 1.16. Outline of the N-terminal COFRADIC strategy.
All cysteine residues are reduced and alkylated. Then, all free amines (α and ε-amines) are acetylated and the protein mixture is proteolysed with trypsin. This creates two types of peptides: N-terminal peptides with blocked N-termini and internal peptides with free N-termini. Following the primary reverse phase HPLC separation of the peptide mixture, peptides present in each fraction are treated with TNBS which modified internal peptides at the α-amino group. During the secondary chromatographic run, which is identical to the primary run, the TNBS modified peptides shift out of their original elution profile. The internal peptides are discarded and the N-terminal peptides, which elute at the same time interval as the primary run are collected and analysed.

Positional proteomics

A novel approach to the selective isolation and recovery of the N-terminal most peptide of each protein in a proteome was also developed in our laboratory, in a proteome simplification strategy termed "positional proteomics" (McDonald *et al.,* 2005). The strategy is as follows: (1) acetylation of a complex protein mixture; (2) digestion of the modified proteins; (3) biotinylation of newly formed amino groups on internal peptides; (4) Removal of biotinylated peptides by streptavidin.

The complex protein mixture is acetylated in its native state, without reduction and alkylation of cysteines. Disulphide bonds are rare in cytosolic proteins, since the cytosol is generally a reducing environment (Derman and Beckwith, 1991). In order to limit the number of analytical steps and simplify the chemistry involved, reduction and alkylation steps are omitted from this method. Subsequent proteolysis reveals new amino groups for all peptides but the N-terminal most peptide. The unwanted internal peptides can then be targeted through the newly exposed amino group by biotinylation and removed by streptavidin Sepharose. The residual positional signature peptides (NTpeps) are then analysed, in the knowledge that each peptide represents a different protein species and is anchored informatically at the N-terminus of the protein (Figure 1.17).

Protein sequence tags (PSTs)

An alternative method for the sorting and isolation of N-terminal peptides has been developed by Kuhn *et al.* This strategy employs a set of chemistries to specifically characterise a group of hydrophobic proteins. The method, referred to as the PST process, (Kuhn *et al.,* 2003; Kuhn *et al.,* 2005) consists of the following steps: (1) solubilisation of the hydrophobic protein mixture using CNBr; (2) reduction and alkylation of the polypeptide mixture; (3) derivatisation of free amino groups using a basic mass tag (BMT); (4) digestion of modified polypeptides; (5) selection of N-terminal CNBr fragments using NHS-activated medium.

When working with hydrophobic proteins it is necessary to incorporate a pre-solubilisation step. In this strategy, cyanogen bromide (CNBr) is used to fragment the insoluble proteins. This stage results in the formation of multiple polypeptide chains for each protein molecule in the system. As a consequence of this step, each polypeptide chain generated by CNBr cleavage will result in a new α-amino group, leading to multiple N-termini for each protein in the mixture. This process was initially applied to a mitochondrial protein suspension from yeast, as follows:

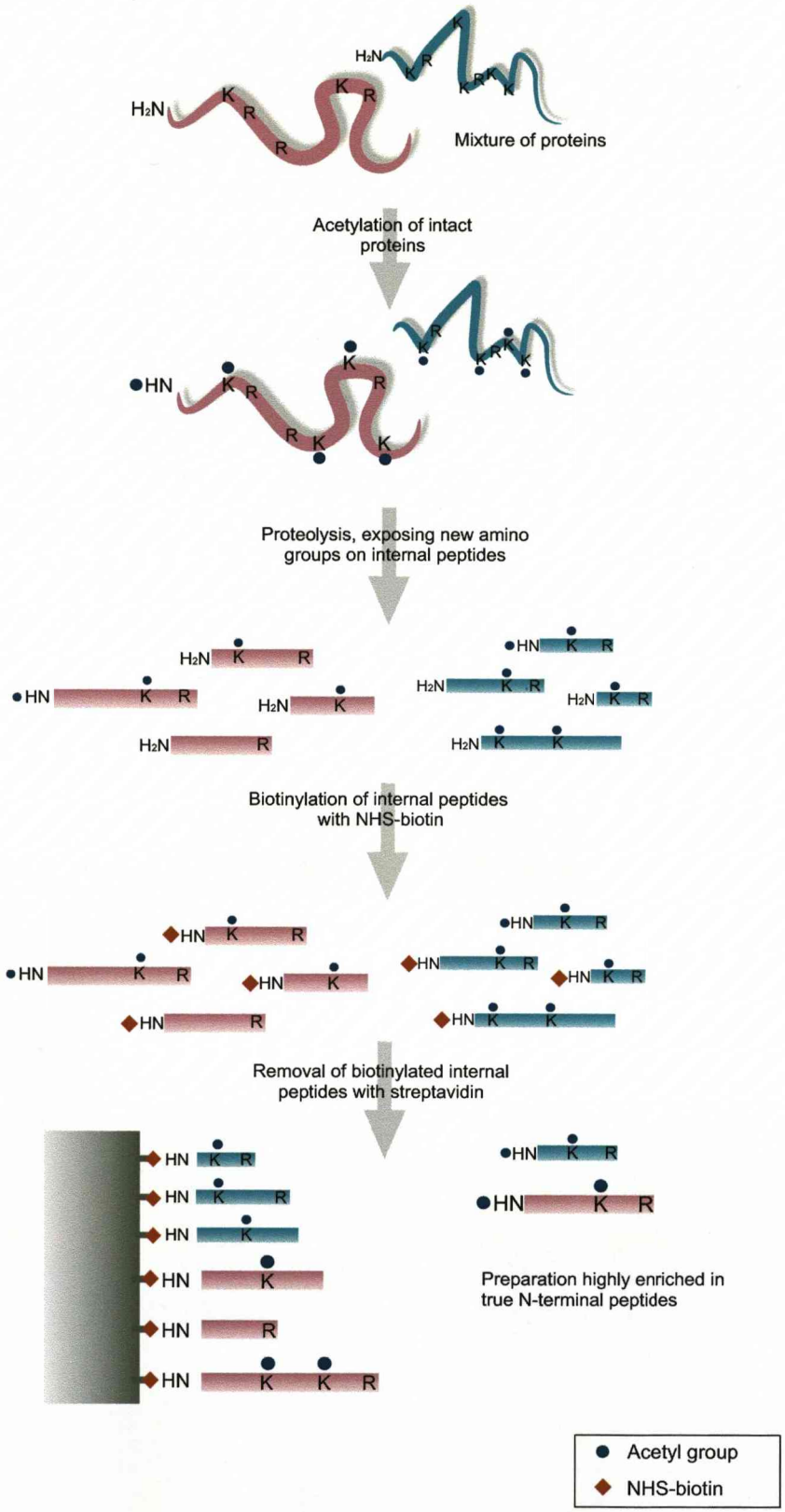


Figure 1.17. Outline of the N-terminal 'positional proteomics' strategy.
Free amino groups (α and ε) are acetylated prior to proteolysis, which results in a mixture of N-terminally acetylated (true N-terminal) and non-acetylated (internal) peptides. Biotinylation of the proteotytically exposed α-amino groups, using NHS-biotin, and subsequent incubation with immobilised streptavidin creates a preparation enriched in N-terminal peptides, which are blocked by acetylation and therefore resistant to biotinylation.

The isolated protein suspension was solubilised by dilution in formic acid followed by addition of CNBr and purification by size exclusion chromatography. The resulting polypeptide mixture was then reduced in order to break disulphide bridges and cysteine residues blocked by alkylation. The free amino groups were then derivatised with BMT which consisted of N,N-dimethylglycine N-hydroxysuccimide ester. To remove all excess labelling reagents, the polypeptide mixture was subjected to a second fractionation step by size exclusion chromatography. The purified mixture was then digested with trypsin, which as in the previously described N-terminal strategies, generated a set of arginine terminated tryptic peptides, due to blockage of lysine residues. The N-terminal sorting stage involved the use of scavenging beads which were prepared using N-hydroxysuccinimide to capture all free amino groups present on the internal peptides. The end result of this process is a pool of peptides that represents the N-terminal most peptide from each CNBr cleavage product. The PST rationale is illustrated in (Figure 1.18).

The use of BMT reagents to block amino groups on amino acid side chains has an additional function. BMTs provide an easily protonatable group at the α-amino group, which has been shown to improve MS/MS fragmentation by enhancing the formation of b-ions (Cárdenas, 1997).

To demonstrate the validity of the PST approach to study membrane proteins, the method was applied to a well characterised biological system, the yeast mitochondrion. This subcellular fraction is known to contain approximately 770 proteins, of which 18% are hydrophobic (Kumar *et al.*, 2002). The PST process resulted in the identification of 147 proteins, including 50 membrane proteins. To compare the PST technology to previously applied gel-based methods, the group also performed a separation of the mitochondrial sample by 2-D SDS-PAGE. This analysis produced 501 gel spots which were subjected to MS analysis, resulting in the identification of 412 proteins corresponding to 112 different gene products. The total number of membrane proteins identified by the 2-D SDS-PAGE approach was 13. This comparison highlights the benefits associated in the application of this method to a set of hydrophobic proteins.

The initial use of CNBr to facilitate solubilisation of hydrophobic proteins results in the generation of multiple polypeptide chains for each protein. A consequence of this step is that each protein will be represented by multiple peptides originating from the newly formed α-amino groups from the CNBr cleaved products. Although this has the advantage of improving coverage of abundant proteins, from a global proteomics perspective, it is not desirable for

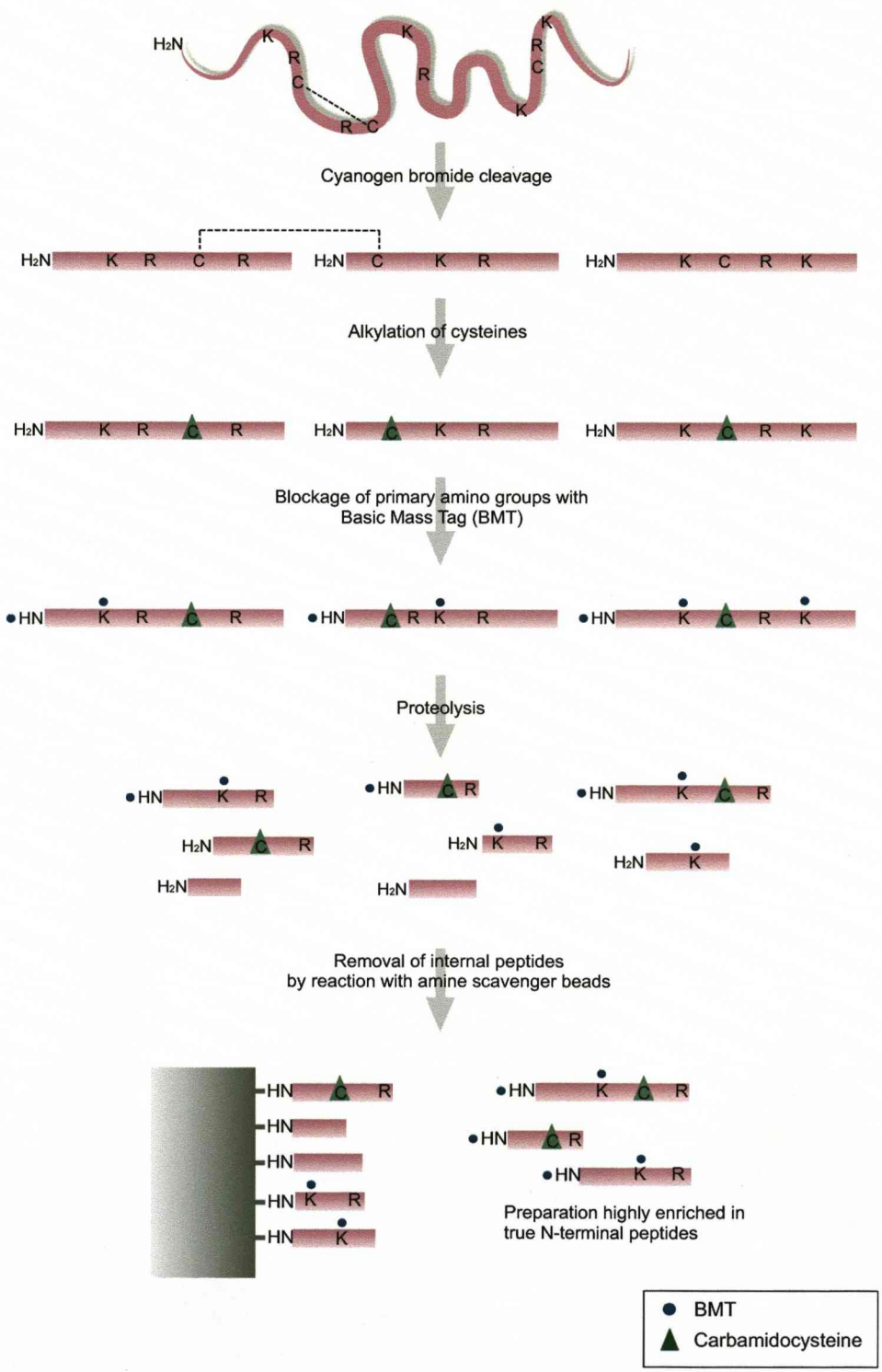


Figure 1.18. Outline of the N-terminal positional sequence tag (PST) strategy.
The first step in the PST isolation process involves chemical cleavage of the protein mixture with cyanogen bromide. This step is necessary in order to solubilise hydrophobic proteins. The polypeptide chains are then reduced and alkylated to break disulphide bonds. Free amino groups (lysine side chains and N-terminal) are blocked using the amine reactive BMT. Proteolysis results in a peptide mixture consisting of BMT blocked N-terminal peptides, from the cyanogen bromide fragments and unblocked internal peptides. Coupling of internal peptides to amine scavenger beads allows separation from the N-terminal peptides which remain in the supernatant.

each protein to be represented by multiple peptides. For this reason the PST procedure is not the best method for global proteomics analysis.

Biotin-avidin method

Another approach for selection of N-terminal peptides has been described by Yamaguchi *et al.* This approach, based on the biotin-avidin strategy utilises a novel biotinylation reagent, biotinylcysteic acid which introduces a sulfonic acid group to the α-amino group on the N-terminal peptide (Yamaguchi *et al.,* 2005). The introduction of a sulfonic group is an effective method to facilitate *de novo* sequencing (Marekov, 2003). Incorporation of the sulfonic group, in addition to the biotin tag, provides a combined method for improved sequencing efficiency and N-terminal isolation. The key steps are as follows: (1) reduction and alkylation of the polypeptide mixture; (2) guanidination of lysine residues (3) biotinylation of the α-amino group; (4) proteolysis of the modified proteins; (5) selection of N-terminal peptides using avidin resin. The full scheme for this procedure is shown in (Figure 1.19).

The guanidination reaction is necessary to prevent biotinylation of lysine side chains at a later stage of the protocol. In this reaction O-methylisourea hemisulfate and NH_4OH, are used to convert lysine residues into homoarginine (Warwood *et al.*, 2006). Following the guanidination reaction the protein mixture is reacted with biotinylcysteic acid which specifically targets the α-amino group of the proteins. Proteolysis generates a peptide mixture consisting of biotinylated N-terminal peptides and unmodified internal and C-terminal peptides. Subsequent incubation of the peptide mixture with avidin resin results in binding of N–terminal peptides, which are eluted from the resin using acetic acid and ACN. The method was validated using the model proteins bovine serum albumin (BSA) and chicken egg white lysozyme. The published protocol was not applied to a complex biological sample.

An adaptation to this method has recently been published in which proteins are initially pre-fractionated by 1-D/2-D SDS-PAGE (Yamaguchi, 2007). The initial stages of the N-terminal sorting procedure were carried out in-gel as follows: The spot of interest was excised from a Coomassie stained 1-D or 2-D SDS gel. Gel spots were washed and dehydrated using ACN. The dried gel piece was then incubated with dithiothreitol to reduce disulphide bridges and iodoacetamide to block cysteine residues. The gel piece was then washed to remove excess reagents and dehydrated for a second time with ACN. The dried gel piece was incubated with the guanidination reagents (as before). Following guanidination the gel piece was washed and dehydrated for the third time. The next stage was biotinylation of the free α-

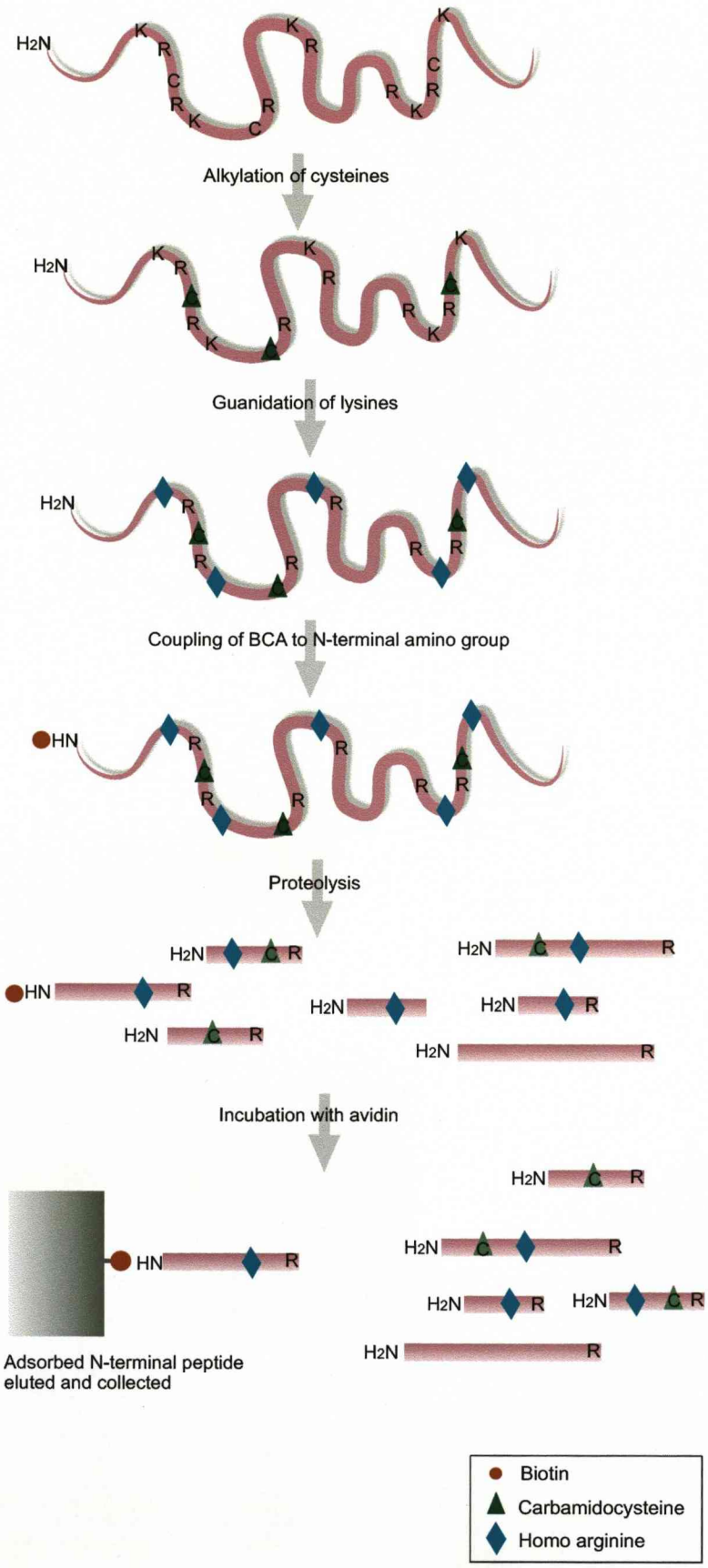


Figure 1.19. Outline of the biotin-avidin method for N-terminal peptide isolation.
All cysteine residues are reduced and alkylated. Then, lysine residues are converted to homoarginine by guanidination. The α-amino group is then derivatised with biotincysteic acid and the protein digested with trypsin. The resulting peptide mixture is passed through an avidin column. The internal peptides pass through the column and are discarded. The biotinylated N-terminal peptides, which are retained by the column, are eluted and subjected to further analysis.

amino group on the N-terminal of the protein. The reagent used for this purpose was sulfo-NHS-SS-biotin which consists of a biotin molecule coupled to a NHS ester reactive group. Following a final wash and dehydration step the gel piece was incubated with trypsin to digest the modified protein. The recovered solution consisted of a mixture of guanidinated internal peptides and guanidinated, biotinylated N-terminal peptides. The peptide solution was then added to a fresh suspension of avidin in which the biotinylated N-terminal peptides bound with high affinity. The internal peptides were washed off the bead suspension and the N-terminal fragments desorbed using ACN and formic acid.

Following optimisation of this protocol using model proteins (BSA, lysozyme and cytochrome C), the method was applied to proteins present in *E. coli* extracts. The cell lysate was first resolved using 2-D SDS-PAGE and stained with Coomassie. A total of 34 protein spots were selected from the gel and subjected to the N-terminal isolation protocol. The resulting N-terminal fragments were analysed by MALDI-ToF MS. This study demonstrates a potential advantage of the method, which is its ability to discriminate between multiple proteins present in the same spot. A major disadvantage of this strategy is the amount of derivatisation and subsequent washing steps involved in the protocol, for this reason this method would not be suitable for global analysis.

Isocyanate resin

The N-terminal peptide purification methods described so far have all required the modification of ε-amino groups present on the side chains of lysine residues. This step is necessary to prevent loss of N-terminal containing lysine residues during the final sorting stage in the protocol. A recent strategy by Mikami *et al.* for the isolation of blocked N-terminal peptides using isocyanate resin, achieves purification without lysine modification (Mikami and Takao, 2007). This method works by exploiting the slight difference in pKa between α-amino (9.06) and ε-amino (10.4) groups. By optimising the reaction conditions to favor coupling of α-amino groups, they have demonstrated it is feasible to isolate *in-vivo* N^{α}-acetylated, lysine containing peptides.

The method can be summarised in four steps: (1) proteolysis of the protein mixture; (2) incubation of the peptide mixture with isocyanate resin; (3) centrifugation and collection of supernatant; (4) MS characterisation of purified blocked N-terminal peptides. The key step in this procedure is coupling of peptides to an isocyanate resin via reaction of the α-amino group on the internal peptides. The isocyanate-coupled resin was prepared by reacting a divalent

isocyanate, methylenediphenyl 4,4'-diisocyanate (MDI), with aminopropyl resin (porous-NH), as a support medium. Under weak acidic conditions, the isocyanate resin specifically reacts with α-amino groups on internal peptides and not with ε-amino groups of lysine. Optimisation of the reaction conditions revealed that a pH lower than 3 resulted in incomplete coupling of α-amino groups to the resin. The optimum pH for the coupling reaction is pH 3.5 – 4.0. Additionally, reaction conditions containing >20% water cause the isocyanate to decompose. Reactions were carried out in 85-90% ACN which also reduced non-specific hydroadsorption of peptides to the resin. Excess use of the resin was shown to lead to non-specific coupling with ε-amino groups of lysine and therefore led to a reduced yield of lysine containing blocked N-terminal peptides (Figure 1.20).

The resin was validated using the synthetic peptide (YGGFLSYPLK), in both its unblocked and N^{α}-acetylated form. Following incubation of the two peptides with the resin in 0.05 M phosphoric acid (pH3.5) in non-aqueous conditions the signal corresponding to the unblocked peptide was no longer present, while the signal corresponding to the blocked peptide was distinctly observed. This experiment suggested that selective capture of unblocked peptides with the isocyanate resin permits exclusive isolation of N^{α}-acetylated peptides without the need for protection of ε-amino groups. A potential disadvantage to this N-terminal isolation strategy is the requirement for a constant low pH. Failure to maintain the required reaction environment would result in loss of lysine containing N-terminal peptides. Although this strategy has the potential to provide a high throughput survey of N-terminal peptides in a complex biological sample, this study did not demonstrate the application of this method to a global proteomic analysis.

Ion exchange chromatography

It has been established that *in vivo* acetylated tryptic peptides originating from the true N-terminus of proteins can be isolated using ion exchange chromatography (Betancourt *et al.*, 2001). This method for N-terminal isolation exploits the reduced basicity of N^{α}-acetylated peptides. In contrast to the majority of N-terminal sorting methodologies this protocol requires no derivatisation steps. Instead, the tryptic peptide mixture from the blocked protein is incubated with carboxypeptidase B which specifically removes the basic amino acid (lysine or arginine) from the C-terminal end of the peptide. At acidic pH, the resulting mixture consists of charged internal peptides and the uncharged N-terminal peptide. The N-terminal peptide is retrieved from the mixture using an SCX resin which is deposited as a thin layer over the top

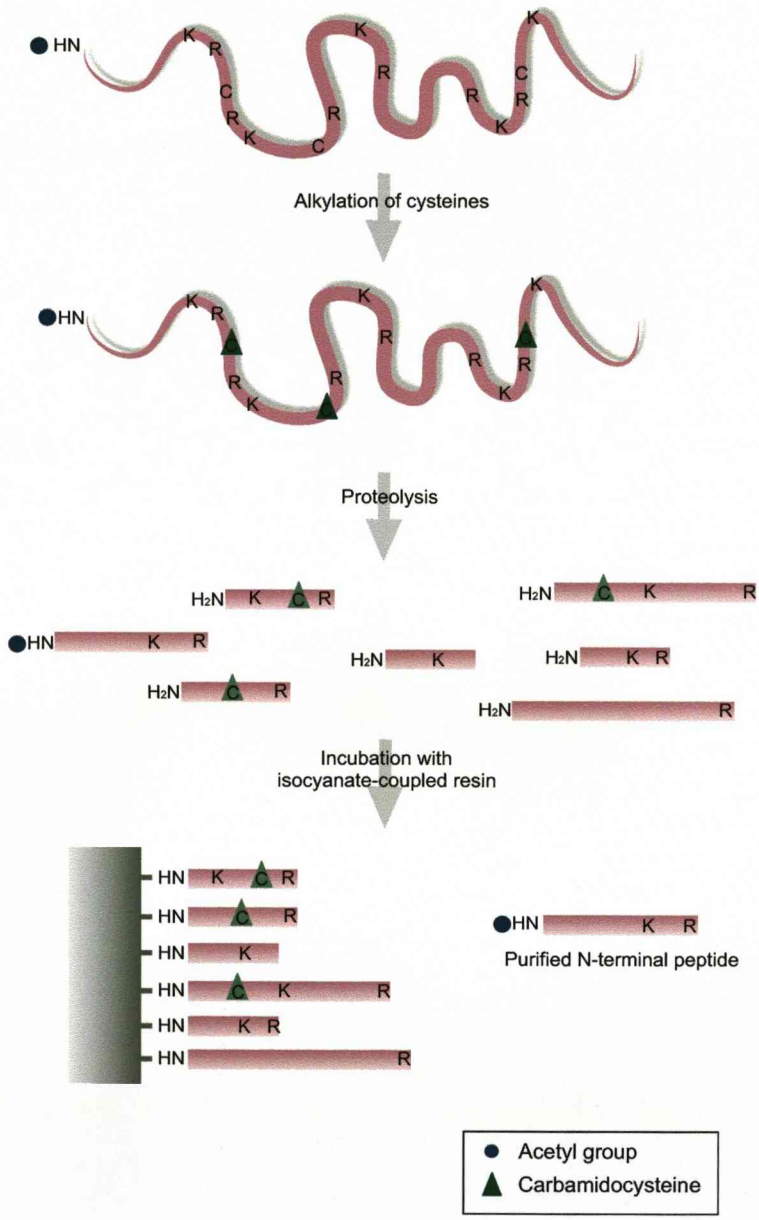


Figure 1.20. Purification of N-blocked peptides using isocyanate resin.
Cysteine residues are first reduced and alkylated. The protein is then proteolysed and the resulting peptides incubated with isocyanate-coupled resin under weak acidic conditions. All internal peptides bind to the resin through interaction with the α-amino group. The blocked N-terminal peptides remain in the supernatant.

of an ultrafiltration tube. Following a brief centrifugation step the N^{α}-acetylated peptide, which is not retained by the resin, is collected for analysis by mass spectrometry.

A more recent method for the targeted analysis of protein termini adopts a similar principle to this ion exchange simplification strategy. Because C-terminal regions of tryptic peptides tend to be basic (arginine or lysine), they also elute later from a SCX column compared to C-terminal and N^{α}-acetylated peptides. This strategy has been applied to the enrichment of N and C-terminal tryptic peptides from a membrane enriched fraction of human embryonic carcinoma cells (Dormeyer *et al.*, 2007). The cells were initially digested using Lys-C and trypsin to yield a complex mixture of lysine and arginine terminated peptides. The rationale behind the use of two enzymes is to broaden the range of peptide length, leading to increased coverage by MS. The peptide mixtures were fractionated by SCX. An initial screen of the fractions by LC-MS/MS analysis revealed that the majority of blocked N-terminal and C-terminal peptides eluted from the SCX column between six and seven minutes, with later fractions mainly comprising internal peptides. This process is summarised in Figure 1.21.

1.9.9 Summary of N-terminal isolation strategies

The N-terminal isolation methods described above are all unique with regard to the chemistries involved and the sorting strategies used. It is therefore important to consider the individual strengths and weaknesses of each protocol before deciding on the most appropriate technique to use. Table 1.1 provides a summary of the methods discussed and highlights their specific advantages and disadvantages.

Identification of a protein from a single peptide sequence has been described in the literature as a 'one hit wonder' (Veenstra *et al.*, 2004), whereby a protein is identified using data derived from a single peptide. If this single peptide was identified using a standard shotgun approach then this would be cause for concern. Part of this concern relates to the lack of information regarding the location of the peptide within the parent protein, and also to the search space required for identification (the entire database of candidate proteins must be given equal validity in the analysis). Techniques that isolate a single specific peptide from complex mixture overcome the issues associated with 'one hit wonders', by anchoring the peptides at a precise location within the parent protein, it is therefore possible to limit the database search to a small subset of peptides.

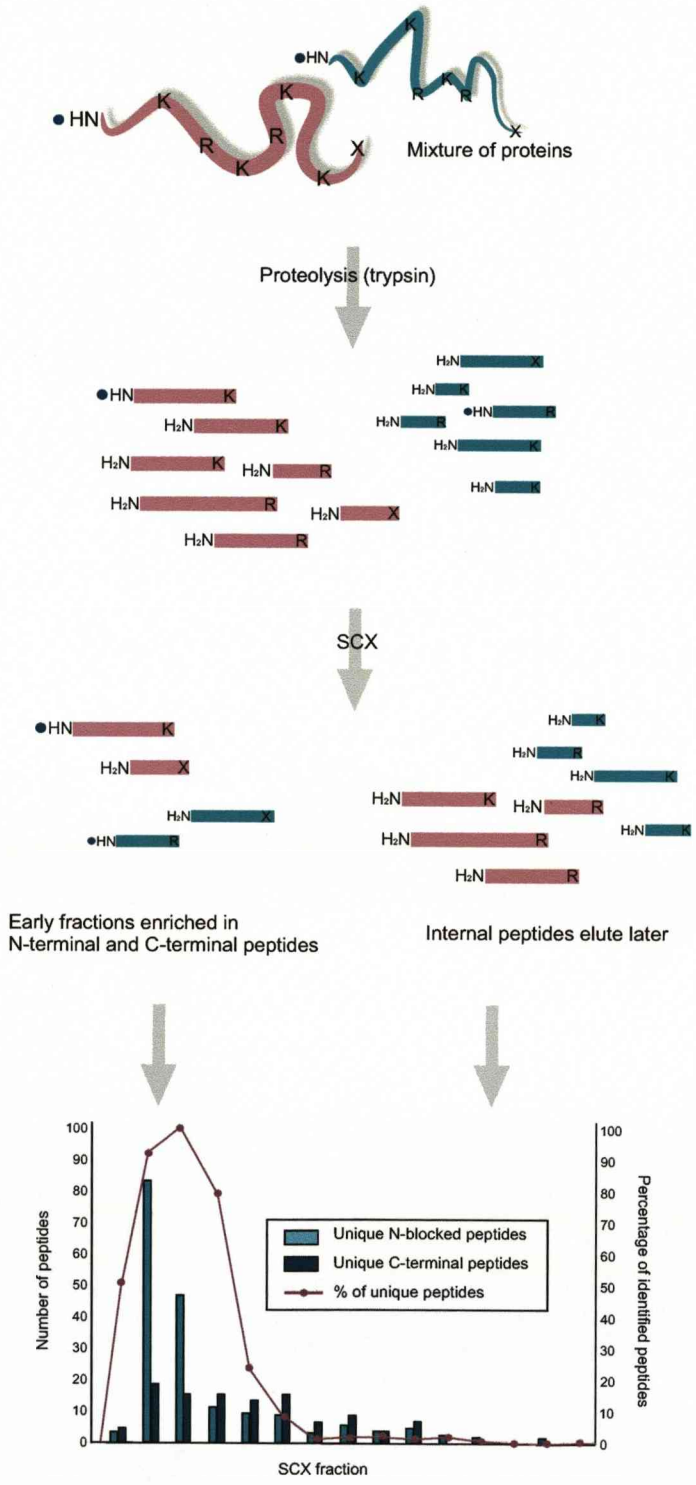


Figure 1.21. Strategy for the enrichment of C-terminal and blocked N-terminal peptides by SCX chromatography.
A mixture of N-terminally blocked (N-acetylated) proteins are proteolysed with trypsin generating a mixture of blocked N-terminal peptides, internal peptides (lysine and arginine terminated) and C-terminal peptides. The entire peptide mixture is then subjected to SCX chromatography. Due to their reduced basicity compared with internal peptides, the blocked N-terminal and C-terminal peptides are enriched in the early fractions and internal peptides are eluted at later intervals. Fractions containing N-terminal and C-terminal peptides are retained for further analysis.

Reference	Key steps prior to MS analysis	Advantages	Disadvantages	*Global*
Gavaert *et al.*	1. Derivatisation of amino groups 2. Proteolysis 3. Primary HPLC run 4. Peptide modification 5. Secondary HPLC run	• Based on a well documented technique (COFRADIC) • Proteome-wide application	• Requirement of multiple chromatographic separations	*Yes*
McDonald *et al.* *	1. Derivatisation of amino groups 2. Proteolysis 3. Biotinylation of peptides 4. Avidin purification	• Internal peptides removed and retained on avidin • Proteome-wide application	• Cysteine peptides not identified	*Yes*
Kuhn *et al.*	1. CNBr cleavage 2. Derivatisation of amino groups 3. Proteolysis 4. Sorting with immobilised NHS	• Suitable for hydrophobic proteins • Internal peptides removed and retained on resin	• More than one peptide isolated per protein	*No*
Yamaguchi *et al.*	1. Guadination of lysine residues 2. Biotinylation of amino groups 3. Proteolysis 4. Avidin purification	• Biotinylcysteic acid facilitates *de novo* sequencing	• N-terminal peptides retained on avidin resin (desorption step required) • Not high-throughput	*No*
Mikami *et al.*	1. Proteolysis 2. Removal of internal peptides with immobilised reagent	• No derivatisation steps	• Requires precise reaction conditions • Only *in vivo* N-acetylated proteins identified	*Yes*
Dormeyer *et al.*	1. Proteolysis 2. Ion exchange chromatography	• No derivatisation steps	• C-terminal peptides also isolated	*Yes*

* Although this thesis describes the development of our N-terminal strategy, it is important to include the method in the overall comparison of current N-terminal isolation methods available

Table 1.2. Strategies for the isolation of N-terminal peptides.
There are a variety of published N-terminal purification protocols available in the literature. Each method differs in the sorting chemistries used.

As discussed earlier, the N-terminal region of a protein molecule is targeted by a variety of modification processes. Therefore, strategies that specifically isolate a protein's N-terminal region provide an efficient way of characterising these modifications in a targeted manner. Events such as signal peptide cleavage, exopeptidase activity, PTMs and methionine removal can all be determined using N-terminal isolation. In addition to determining the true N-terminal status of a protein, isolation of a specific peptide from each protein in a mixture provides a powerful means of simplification. The failure of shotgun proteomic techniques to provide a comprehensive analysis of individual proteomes is mainly due to the vast complexity of tryptic digests. Reducing the products of proteolysis to as little as one peptide per protein addresses the issue of complexity and provides an alternative strategy to standard shotgun approaches. The use of positional proteomic strategies in combination with other protein simplification techniques such as MudPIT or Protein Equalizer™ technology could potentially lead to increased depth of proteomics analysis and bring the goal of global proteomics one step closer.

1.10 Aims and objectives

This thesis describes the progressive development of a novel N-terminal terminal enrichment strategy termed "positional proteomics". The effectiveness of this strategy is demonstrated on a variety of complex biological samples in order to achieve efficient and rationalised simplification.

Chapter 2 describes the methodologies used throughout the study including sample preparation and mass spectrometric instrumentation. Chapter 3 explains the concepts behind the N-terminal simplification protocol, including optimisation of the various chemistries involved using model peptides and purified proteins, before application of the strategy to a variety of complex biological samples. Chapter 4 focuses on the use of this method to characterise human plasma and the issues associated with biomarker based applications. Finally, Chapter 5 introduces the rationale behind a novel isotope coded acetylation reagent which is utilised to quantify the occurrence of lysine residues in proteolytic peptides. The approach, termed **M**ass Isotopomer **D**istribution **A**nalysis of amino acid **R**esidues (**MIDAR**), lends itself well to positional proteomics based strategies as it brings with it additional information regarding amino acid composition.

2. MATERIALS AND METHODS

2.1 REAGENTS

General

- HPLC grade water
- HPLC grade acetonitrile (ACN)
- Trichloroacetic acid (TCA)
- Trifluoroacetic acid (TFA)
- Formic acid (FA)
- Diethyl ether ($C_4H_{10}O$)
- Dimethylformamide (DMF)
- Hydrochloric acid (HCl)
- Methanol (MeOH)
- Acetyl chloride (CH_3COCl)

Sample preparation

- Homogenisation buffer: 20mM Na_2HPO_4 (pH7.5)
- Luria broth (LB; Merck, Nottingham, UK)
- Bugbuster protein extraction reagent (Novagen, Nottingham, UK)
- YPD medium : 1% (w/v) yeast extract, 2% (w/v) peptone, 2% (w/v) glucose
- Yeast lysis buffer: 20mM HEPES pH7.4, 100mM KOAc pH7.6, 2.5mM MgAc, 2mM DTT, yeast protease inhibitor cocktail (Sigma)
- Protease inhibitor tablets (Roche Diagnostics, Lewes, UK)
- Coomassie Plus® protein assay (Pierce, Northumberland, UK)
- Protein Equalizer™ beads (ligand library beads; Louisville, KY, USA)

1-D SDS-PAGE

- 2 x SDS-PAGE sample buffer: 62.5mM Tris HCl pH6.8, 10% (v/v) glycerol, 100mM DTT, 2% (w/v) SDS, 0.001% (w/v) bromophenol blue.
- SDS-PAGE running buffer: 25mM Tris, 192mM glycine, 1% (w/v) SDS
- Coomassie brilliant blue stain: 0.25% (w/v) Coomassie brilliant blue, 50% (v/v) methanol, 12.5% (w/v) acetic acid

- Coomassie destain solution: 50% (v/v) methanol, 5% (v/v) acetic acid

In-gel Digestion

- Digestion buffer: 10mM NH_4HCO_3
- Destain solution: 50% (v/v) acetonitrile (ACN), 50% (v/v) 100mM NH_4HCO_3 (pH 8.2)
- Reduction solution: 10mM dithiothreitol (DTT) prepared in 10mM NH_4HCO_3
- Alkylation solution: 55mM iodoacetamide (IAN) prepared in 10mM NH_4HCO_3
- Extraction buffer: 1% (v/v) FA, 2% (v/v) ACN

Desalting (ZipTip) solutions

- Wash buffer: 0.1% (w/v) trifluoroacetic acid (TFA)
- Elution buffer: 50% (v/v) ACN, 0.1% (w/v) TFA

N-terminal purification

- Acetic anhydride (Fisher Scientific, Leicestershire, UK)
- Acetic anhydride D6 (Cambridge Isotope Laboratories, Andover, MA, USA)
- Streptavidin Sepharose™, High Performance (GE Healthcare, Bucks, UK)
- NHS-activated Sepharose™ 4 Fast Flow (GE Healthcare)
- NHS-Biotin (Pierce)
- Sulfo-NHS acetate (Pierce)
- Amine scavenging beads (Tris (2-aminoethyl) amine polymer bound; Sigma, Dorset, UK)
- Acetylation buffer: 20mM sodium carbonate (Na_2C0_3), pH 8.5
- Digestion buffer: 20mM sodium phosphate (Na_2HPO_4), pH7.5
- Binding buffer: 20mM Na_2HPO_4 and 0.15M sodium chloride (NaCl), pH 7.5
- 1mM HCl for NHS-activated Sepharose washing

MIDAR Reagent

- $(C[^2H_3]CO)_2O$ and $([^{13}C]H_3[^{13}C]O)_2O$ mixed in a 9:1 ratio (Cambridge Isotope Laboratories)

MALDI solutions

- MALDI matrix: ~10mg α-cyano-4-hydroxycinnamic acid (CHCA; Sigma) in 1ml, 50% (v/v) ACN, 0.1% (v/v) TFA

Reversed-phase HPLC solutions

- RP running buffer (A) 2% (v/v) ACN:0.1% (v/v) FA
- RP eluting buffer (B) 90% (v/v) ACN:0.1% (v/v) FA

2.2 EQUIPMENT

- MALDI-R, MALDI-ToF mass spectrometer (Waters, Manchester, UK)
- ESI-QToF Micro, tandem mass spectrometer (Waters)
- LTQ ion trap, tandem mass spectrometer (Thermo Scientific, Hemel Hempstead, UK)
- LTQ Orbitrap hybrid MS (Thermo Scientific)
- Ultimate 3000 nanoflow chromatography system (Dionex, Surrey, UK)
- RP column: C18, 3µm particle size (100Å), 75µm diameter × 150mm long (LC Packings, Dionex)
- Multiskan plate reader (Thermo Scientific)
- Slide-A-Lyzer® dialysis cassettes, 500µl to 3ml, 10,000 mw cut-off (Pierce)
- ZipTip™ C18 RP pipette tips (Millipore, Watford, UK)
- pH paper (Sigma)
- Homogeniser (Ystral; SIC, Hampshire, UK)
- Vacuum centrifuge (Jouan RC10.22; Thermo Scientific)

2.3 SOFTWARE

- MassLynx version 4 (Waters)
- Xcalibur version 2.0.7 (Thermo Scientific)
- Chromeleon version 6.7 (Dionex)
- Ascent software (Thermo Scientific)

2.4 Samples

In this study, a range of model peptides, purified proteins and biological samples were used for method development and proteomic analysis.

Model peptides

Des-Arg-bradykinin, neurotensin, adrenocorticotrophic hormone (ACTH) fragment 1-17, ACTH fragment 18-39, [Glu1]-Fibrinopeptide B and insulin β-chain were all purchased from Sigma. Details, including sequence and $[M+H]^+$ values can be found in Table 2.1.

Model proteins

Bovine serum albumin (BSA; Pierce), triose phosphate isomerise (TPI) from rabbit muscle and pyruvate kinase (PK) from rabbit muscle (Sigma). Tryptic and Arg-C peptide maps are displayed in Figure 2.1.

Mouse skeletal muscle and liver soluble protein fractions

Skeletal muscle and liver tissue from the house mouse (*Mus musculus domesticus)* were obtained in-house and dissected immediately post-mortem. The tissue (1g wet weight) was homogenised in 20mM Na_2HPO_4, pH7.5 (10ml) containing protease inhibitors. The whole homogenate was centrifuged for 45min at 13,000 *x g* at 4°C. The resultant supernatant fraction was stored at -20°C or used immediately without further purification. The concentration of the soluble fraction was determined using the Coomassie Plus Protein assay.

E. coli cell lysate

A single colony of *E. coli* strain BL21(λ)DE3, of genotype: F-, ompT, $hsdS_B$ ($r_B^-m_B^-$) gal,dcm was used to inoculate 10ml Luria broth (LB) and incubated overnight at 37 °C with shaking. The overnight culture (0.5ml) was inoculated into 50ml of fresh LB media, prewarmed to 37°C (1:100 dilution) and incubated with shaking. Samples were removed at hourly intervals and the absorbance at 600nm determined. Growth rate was monitored until early stationary phase was reached. The culture was transferred to a pre-weighed 50ml centrifuge tube and centrifuged at 1,200 *x g* for 10min at 4 °C. The supernatant was decanted and the tube weighed again to determine the wet weight of the cell pellet. For 1g of wet cell pellet 2.5ml Bugbuster protein extraction reagent was added and the cells were placed on a rocker platform at room temperature for 20min in order to ensure good resuspension and cell breakage. Workflow to this point was conducted by Deborah Simpson. The supernatant was removed (soluble

	Peptide	$[M+H]^+$	Sequence
1	Des-Arg-bradykinin	904.47	RPPGFSPF
2	Neurotensin	1072.92	pELYENKPRRPYIL
3	[Glu1]-Fibrinopeptide B human	1571.20	EGVNDNEEGFFSAR
4	ACTH, human, rat fragment 1-17	2093.09	SYSMEHFRWGKPVGKKR
5	ACTH, human, rat fragment 18-39	2465.20	RPVKVYPNGAEDESAEAFPLEF
6	Insulin β-chain	3495.89	FVNQHLCGSHLVEALYLVCGERGFFYTPKA

(pE = pyroglutamate)

Table 2.1. Sequences and $[M+H]^+$ values of model peptides used throughout this thesis.

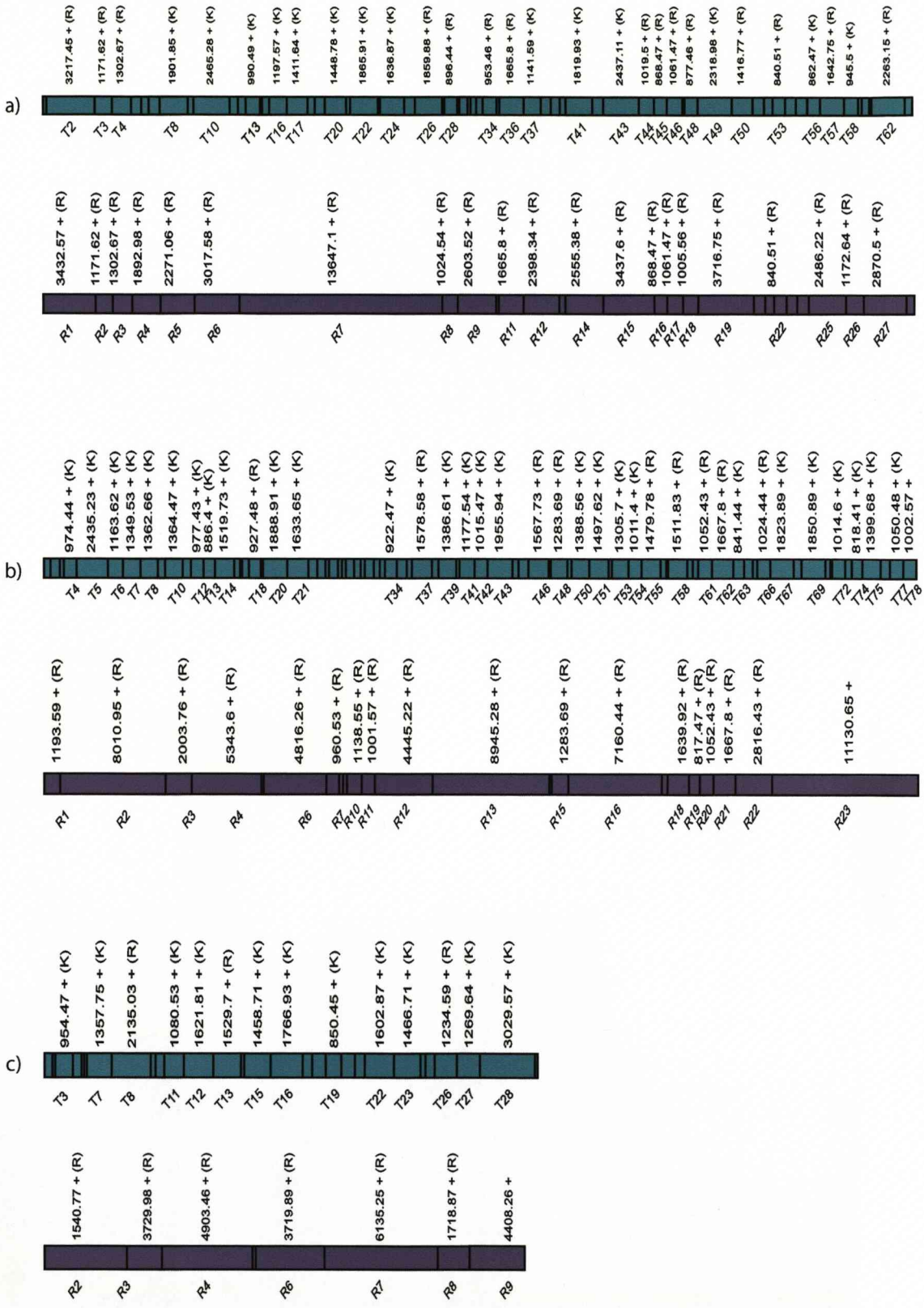


Figure 2.1. Peptide maps for purified proteins.

Amino acid sequences for rabbit pyruvate kinase (a), bovine serum albumin (b) and rabbit triosephosphate isomerase (c) were retrieved from SwissProt, imported into Peptide Mapper and subjected to *in silico* proteolysis using trypsin (green) and Arg-C (purple). Monoisotopic masses $[M+H]^+$ and C-terminal amino acid residue are labelled for each peptide

fraction) and used immediately or stored at -20 °C. The protein concentration was determined using the Coomassie Plus Protein assay.

The soluble fraction of the *E. coli* cell lysate (3 ml) was dialysed into 20mM Na_2CO_3 pH 8, using Slide-A-Lyser cassettes ® . The sample was dialysed for 4h at room temperature, with stirring. The concentration of the dialysed sample was determined using the Coomassie Plus Protein assay.

S. cerevisiae cell lysate

Cell extracts were provided by John Hughes and Abigail Stevenson of the McCarthy Lab (University of Manchester). *S. cerevisiae* cells ("wild-type" strain PTC49) were grown in YPD medium in shaking incubators (200 rpm) at 30°C, harvested and cell lysates prepared by bead-beating in yeast lysis buffer.

Human plasma samples

Pooled human plasma was obtained lysophilised from Sigma. Individual samples were obtained from healthy volunteers. Plasma was extracted by centrifugation from heparinised peripheral venous blood samples.

Resolving gel

	Reagent	Amount
1	30% (w/v) acrylamide	4.0 ml
2	1.5M TrisHCl pH 8.8	2.5 ml
3	MilliQ water	3.3 ml
4	10% (w/v) SDS	50.0 µl
5	10% (w/v) APS	75.0 µl
6	TEMED	7.5 µl

Stacking gel

	Reagent	Amount
1	30% (w/v) acrylamide	0.6 ml
2	0.5M TrisHCl pH 6.8	1.25 ml
3	MilliQ water	3.07 ml
4	10% (w/v) SDS	50.0 µl
5	10% (w/v) APS	25.0 µl
6	TEMED	2.5 µl

Table 2.2. Resolving and stacking gel solutions for 1-D SDS-PAGE.
Quantities are for two mini (10cm) acrylamide gels (10ml resolving gel/5ml stacking gel).

2.5 General protocols

2.5.1 1-D SDS-PAGE

SDS-PAGE separations were performed according to manufacturers instructions. Recipes for gel solutions are described in Table 2.2. Samples were diluted (1:1) by the addition of an equal volume of 2 x SDS-PAGE sample buffer and heated at 100°C for 5min. Samples were allowed to cool and centrifuged briefly. The appropriate volume of sample was loaded onto the gel to deliver 10-15µg protein. Gels were placed in Coomassie brilliant blue stain overnight with rocking. To destain, gels were placed in Coomassie destain solution.

2.5.2 In-gel proteolysis

Gel plugs containing protein spots of interest were excised from 1-D gels using a glass pipette and transferred to a 96-well plate. To each tube 25µl of destain solution was added and incubated at 37 °C for 20min. This process was repeated until all of the stain had been removed. The plugs were then washed in 50mM ammonium bicarbonate, which was subsequently discarded. Reduction solution (25µl) was added to each plug and incubated at 37 °C. After 30 min, the supernatant was discarded, alkylation solution (25µl) was added to each tube, and incubation continued in the dark for 60min. The gel was dehydrated using 25µl of ACN, and incubation at 37°C was resumed for 15min. The supernatant was removed from the dehydrated plug, which was allowed to air dry. Once dry the gel was rehydrated in digestion buffer (9µl) containing trypsin (Roche Diagnostics; 1µl of 100ng/µl trypsin stock reconstituted in 50mM acetic acid). After 30min, 50mM NH_4HCO_3 (10µl) was added to each tube, and digestion was allowed to continue overnight at 30 °C.

2.5.3 Esterification of peptides

A stock solution of MeOH (1ml previously stored at -20 ° C for 15min) and acetyl chloride (150µl) was prepared immediately prior to use. An aliquot (10µl) of this mixture was added to a portion of desalted tryptic peptides (purified using ZipTip™ C18 pipette tip), previously dried under vacuum. The mixture was incubated at room temperature for 45min prior to drying in a vacuum centrifuge. Esterified peptides were analysed by MALDI-ToF MS.

2.5.4 In-solution proteolysis

Proteins were reduced by the addition of DTT (50mM) and alkylated by the addition of IAN (100mM). Proteins were precipitated by adding five volumes of cold TCA (30% (w/v)), samples were vortexed and incubated on ice for 1h. Samples were then centrifuged at 13,000 *x g* for 2min to pellet the protein. The TCA was removed from the tube and discarded. Diethyl ether (200µl) was added to the protein and agitated using a pipette tip. The sample was centrifuged at 13,000 *x g* for 10sec in order to re-pellet the protein. The pellet was washed in diethyl ether a total of three times, in order to completely remove residual TCA. The tube containing the pellet was placed at 37°C for 5min with the lid open to ensure removal of ether. Once dry the pellet was resuspended in 50mM digestion buffer (10µl) containing trypsin (1µl of 100ng/µl trypsin stock reconstituted in 50mM acetic acid). Digestion was allowed to continue overnight at 37°C.

2.6 Development of the N-terminal isolation strategy

2.6.1 Acetylation of proteins

Acetylation reactions were performed on 50µg (~30nmol) of the soluble protein mixtures described above. The material was treated with 1mg (30µmol) sulfo-NHS acetate or 1µl (10µmol) acetic anhydride (~1000-fold excess of reagent), for 60min at pH9. Following this reaction, excess reagent was removed using a 10-fold excess (10mg) of amine scavenging beads (binding capacity 3.5-5.0mmol/g), a treatment that had a major effect on the overall success of the process, and obviated the addition of free amines to inactivate excess reagent.

2.6.3 Proteolysis of acetylated proteins

Following TCA precipitation (as above), the protein pellet was resuspended in 20mM Na_2HPO_4, pH7.5 (50µl) and digested overnight at 37°C with 1µg trypsin (1:50 enzyme:substrate).

2.6.4 N-terminal recovery using biotin/streptavidin method

Biotinylation was performed on 1µg (~0.6nmol) of desalted (C18 ZipTip), acetylated peptide mixture, reconstituted in binding buffer (20mM Na_2HPO_4, 0.15M NaCl, pH 7.5; 20µl). The peptides were modified with a 100-fold molar excess NHS-biotin, prepared immediately before use by reconstituting 1mg of biotin in 100µl DMF and 2µl (60nmol) of biotin was added to the desalted peptide mixture. The biotinylation reaction was allowed to proceed for 2h at room temperature. Excess biotin was removed from the sample by the addition of a 100-fold excess of amine scavenging beads (1mg). Biotinylated peptides (1µg) were desalted on a C18 ZipTip followed by the separation of biotinylated internal and unbiotinylated N-terminal peptides on streptavidin Sepharose (High Performance). The Sepharose (20µl) was washed three times with binding buffer before adding 1µg (~0.6 nmoles) of desalted peptides (made up to 20µl in binding buffer). The flow through was retained and the column was washed with a further 10µl of wash buffer. The unbound material was pooled and analysed by MS without further treatment.

2.6.5 N-terminal recovery using NHS-activated Sepharose

NHS-activated Sepharose™ 4 Fast Flow, stored in propanol (50µl; GE Healthcare), was transferred to a 1.5ml microcentrifuge tube and washed twice with cold 1mM HCl, then once with binding buffer (20mM Na_2HPO_4, 0.15M NaCl, pH 7.5) immediately before use. The total acetylated digest (approximately 30nmol of digested peptides) was made up to 100µl in binding buffer and added to the beads. The Sepharose contains 18µmol NHS/ml of medium, therefore, 50µl of Sepharose will provide 900nmol of NHS, giving a 30-fold excess. The mixture was vortexed briefly then incubated for 4h at room temperature on a rotary mixer. The mixture was centrifuged and the supernatant removed. The supernatant was then added to another 50µl aliquot of NHS-activated Sepharose (equilibrated as before) and the mixture was incubated overnight at 4°C with continuous mixing. The supernatant containing unbound peptides was removed and analysed by MS without further treatment.

2.6.6 Reversal of O-acetylation

Partial acetylation of serine and tyrosine residues was reversed by the addition of 1µl (30 µmol) hydroxylamine to the final N-terminal preparation.

2.7 Normalisation of human plasma using Protein Equalizer™ beads

Lyophilised human plasma (Sigma) was reconstituted in 20mM Na_2CO_3, pH8.5 (1ml), to give a final protein concentration of 30mg/ml (concentration determined using the Coomassie Plus Protein assay). Protein Equalizer™ beads were prepared for use as follows: The beads (20mg) were weighed out and transferred to a 0.5ml microcentrifuge tube. MeOH 100% (v/v; 1ml) was added to the beads and the suspension was incubated for 10min, on a rotary mixer. Beads were allowed to settle and the supernatant was removed and discarded. MeOH 50% (v/v; 1ml) was added to cover to surface of the beads, which were allowed to swell overnight at 4° C. Once swollen, 20mg beads (constituting 100µl settled bead volume) were transferred to a 1.5ml microfuge tube. Beads were washed in 1ml double distilled H_2O on a rotary mixer for 30min prior to equilibration by repeated washing in 20mM Na_2CO_3, pH8.5 for 30min (three washes in total). After each wash, beads were left to settle for 5min and the supernatant removed. Approximately 1ml sample containing 30mg soluble protein was added to the beads and mixed for 2h with turning. The supernatant fraction, containing unbound material, was collected after beads had settled for 5min. The beads were subsequently washed eight times in 1ml 20mM Na_2C0_3 and each supernatant fraction was retained for analysis.

2.8 N-TERMINAL ISOLATION OF PROTEINS BOUND TO PROTEIN EQUALIZER™ BEADS

An adapted N-terminal isolation protocol was performed directly with plasma proteins coupled to Protein Equalizer™ beads, as follows: A portion of beads containing bound plasma proteins (20µl) was removed from the final suspension. The bead/protein mixture was acetylated by the addition of 1mg sulfo-NHS acetate (reconstituted in 30µl acetylation buffer). The acetylation reaction was allowed to proceed at room temperature for 2h. The reaction was stopped by the addition of 50µl 1M Tris-HCl, pH 9, which serves to quench residual acetylation reagent. The beads were allowed to settle and the supernatant was removed. The beads were washed repeatedly in 20mM Na_2HPO_4, pH 7.5, (1ml) a total of five times. Once the final wash had been removed, 20mM Na_2HPO_4 pH 7.5 (20µl) was added to the beads and the bound proteins digested overnight at 37°C with 1µg trypsin (1:50 enzyme:substrate). The digested peptides were removed from the beads in the supernatant and subjected to N-terminal isolation using NHS-activated Sepharose. Immediately prior to use the Sepharose (50µl) was washed twice with cold 1mM HCl then once with binding buffer. The peptide mixture was made up to 50µl in binding buffer and added to Sepharose. The mixture was vortexed briefly then incubated for 4h at room temperature on a rotary mixer. The mixture was centrifuged and the supernatant removed. The filtered peptide containing solution was then added to another 50µl aliquot of NHS-activated Sepharose (equilibrated as before) and the mixture was incubated overnight at 4°C on a rotary mixter. The supernatant containing the unbound N-terminal peptides was removed and analysed by MS without further treatment.

2.9 Chromatography

Peptide mixtures were separated by RP-HPLC prior to ESI-MS analysis. For ESI-LC-MS/MS analysis on the ion trap mass spectrometer peptides were separated using the Ultimate 3000 HPLC system fitted with a C18 RP column (C18, 3µm particle size (100Å), 75µm diameter × 150mm long). Depending on the sample complexity, either the standard 60min gradient or an extended 180min ACN gradient was used (Figure 2.2a and b) with a flow rate of 300nL/min. For analysis on the LTQ-Orbitrap hybrid MS, chromatographic separations were carried out using a surveyor MS pump and Micro AS autosampler (Thermo Fisher Scientific). Chromatographic separation was achieved using a linear gradient of 0% B to 40% B in 160min at a flow rate of 300nl/min (Figure 2.2c).

2.10 Mass Spectrometry

Mass spectrometers used to acquire data in this thesis included, MALDI-ToF MS, ESI-Q-ToF MS, linear quadruple ion trap (LTQ). In addition, selected samples were ran on an LTQ Orbitrap hybrid MS by Gary Woffendin.

2.10.1 ZipTip™ Sample Preparation

The C18 RP ZipTip was prepared by wetting in 10µl of elution buffer (50% (v/v) ACN, 0.1% (w/v) TFA) a total of three times. The tip was equilibrated in wash buffer (0.1% (w/v) TFA) a total of three times prior to sample loading. The sample to be desalted was loaded onto the ZipTip in 10µl aliquots until all the sample had passed through the tip. The ZipTip was then washed using wash buffer (0.1% (w/v) TFA) a total of three times. The desalted peptides were then eluted into 10µl elution buffer for MS analysis.

3.10.2 MALDI-ToF MS Analysis

Peptide preparations were mixed in a 1:1 ratio with matrix solution and spotted directly onto a MALDI target for analysis. Samples were allowed to air dry and analysed using a MALDI-ToF mass spectrometer (M@LDI™; Waters). The mass spectrometer was calibrated using a four point calibration of standard peptides: 12pmol des-arg bradykinin (mw 903.47), 12pmol neurotensin (1,671.92), adrenocorticotrophic hormone (2,464.2) and 75pmol insulin β chain (3,493.65). Data was collected over the range of 900-3500 Thomsons. Instrument operating parameters included: pulse voltage of 3200V, source voltage of 15000V, reflectron voltage of

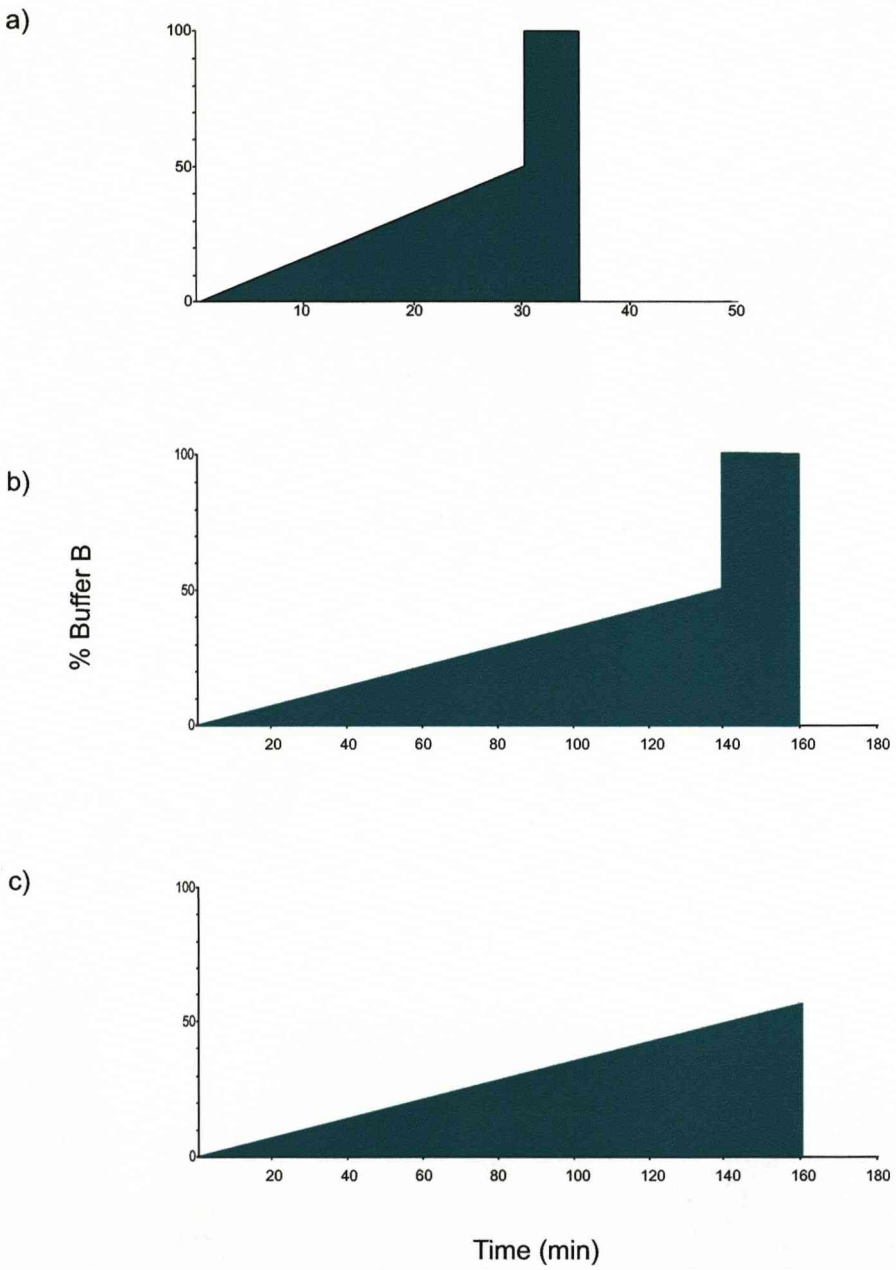


Figure 2.2. Acetonitrile HPLC gradients.
Depending on the complexity of the analyte and the instrument used, a choice of three RP gradients were implemented. For LC-MS/MS on the LTQ ion trap instrument either a standard one hour gradient (a) or an extended three hour gradient was used. For analysis on the LTQ Orbitrap mass spectrometer a linear three hour gradient (c) was implemented.

500V and MCP voltage of 1950V. Matrix suppression was set to 400 mass units. Spectral acquisition was controlled through the MS tune page. Data acquisition and processing was performed through the MassLynx software suite (version 4.0). For spectral acquisition, the laser energy was optimised and acquisition parameters set at 5 laser shots/sec and 10 laser shots/scan. Raw data was processed by combining between 20-30 spectra, subtracting background noise (polynomial order 10 with 40% of the data points below this polynomial curve and a tolerance of 0.01) and smoothing the data (Savitzky Golay method).

2.10.3 ESI Q-ToF MS/MS

Simple peptide mixtures were desalted into 50% (v/v) ACN: 0.1% (v/v) FA using a C18 ZipTip and infused directly into the ESI-Q-ToF micro (Waters) at a flow rate of 0.5μl/min. To scan for precursor ions data was acquired over the range 400-2000m/z with the capillary voltage set at 1900V, collision energy 10V and sample cone 55V. Precursor ion charge states were determined by manual inspection of the isotope envelope, noting the m/z difference between the monoisotopic and the first [^{13}C] peak. For MS/MS analysis, the mass range of the quadrupole was set to transmit the precursor ion and fragment ion spectra collected over the expected mass range, and the peptide was fragmented by CID using a collision energy of 30%. Once the fragmentation pattern of the product ion spectrum was optimised by tuning the collision energy, up to 100 individual spectra were collected and combined to produce to final mass spectrum for processing. The product ion spectrum was processed using MaxENT3 software.

2.10.4 Quadrupole ion trap MS/MS

A linear quadrupole ion trap (LTQ; Thermo Scientific) was used for high throughput protein identification from in-gel digests, in-solution digests and N-terminal preparations. Ionised peptides were analysed in the mass spectrometer, by DDA, using the "triple play" mode, consisting initially of a survey (MS) spectrum from which the three most abundant ions were determined. The charge state of each ion was then assigned from the C13 isotope envelope "zoom scan" and finally subjected to a third MS/MS scan. The LTQ was tuned using a 500fmol/μl solution of GFP and calibrated according to manufacturers instructions.

2.10.5 Orbitrap data acquisition

The LTQ Orbitrap mass spectrometer consists of an LTQ XL linear ion trap mass spectrometer with an Orbitrap high resolution accurate mass detector. The instrument was operated in parallel mode that provided high resolution (set to 30,000) and accurate mass full scan data concurrently with high sensitivity MS/MS peptide fragment ion spectra from the LTQ XL.

2.11 Protein Identification

2.11.1 Peptide mass fingerprinting

The monoisotiopic mass values (m/z) were manually extracted from a single MALDI-ToF mass spectrum and compiled into a list. The list of m/z values was used to search a variety of databases through the Mascot server (Peptide Mass Fingerprint search) against the SwissProt. Search parameters allowed a single missed tryptic cleavage, carbamidomethyl modification of cysteine (fixed) oxidation of methionine (variable) and a peptide tolerance of ±150ppm. The taxonomic space was restricted to the species being searched. Peptide maps, indicating peptide coverage, were produced using the peptide mapping tool (Beynon, 2005).

2.11.2 Manual (*de novo*) sequencing

Processed MS/MS data obtained from direct infusion of peptide mixtures into the ESI-QToF was sequenced (*de novo*) using the PepSeq software in the MassLynx package. The resulting sequence tags were used to search the NCBI database using the BLASTP algorithm (Altschul *et al.,* 1997).

2.11.3 MS/MS ion search

Xcalibur raw data files, generated by LC-ESI-LTQ MS/MS, were processed using Bioworks browser (Thermo Scientific) using the turbosequest tool. DTA (peak list) files for each sample were merged and converted to Mascot Generic Format (MGF) for database searching. MS/MS data were searched using the Mascot search engine (MS/MS ion search) with varying parameters depending on the nature of the sample being searched. In all cases, a peptide tolerance of 1.5Da and an MS/MS tolerance of 0.6Da were used. The data format was set to Mascot generic and the instrument set to ESI-TRAP.

Protein identifications are ranked according to a probability based Mowse score, which is assigned to each peptide ion matched (also known as the ion score; Pappin *et al.*, 1993). Individual ion scores are reported for each peptide matched (from MS/MS data), and these scores are combined to give the protein score. The score for an MS/MS match is based on the absolute probability (P) that the observed match between the experimental data and the database sequence is a random event. For each Mascot search an ion score significance threshold is reported based on -10Log(P). For example, if 1.5×10^5 peptides fell within the mass tolerance window about the precursor mass, and the significance threshold was chosen to be 0.05, (a 1 in 20 chance of a false positive), this would translate into a score threshold of 65.

Searching and identifying proteins from in-solution digests

When searching MS/MS data generated by in-solution, tryptic, proteolysis, the enzyme specificity was set to trypsin and the variable modification of oxidised methionine was chosen. MS/MS data was searched against the SwissProt database and the taxonomy was restricted to the experimental species. Protein matches with scores over the reported threshold indicate extensive homology and, therefore, these were taken as significant identifications.

Strategy for searching and identifying N-terminal peptides

When searching data from N-terminal preparations, the enzyme specificity was set to Arg-C (no cleavage is observed at lysine residues), fixed modifications of acetyl N-term (+42Da at the N-terminus) and acetyl lysine (+42Da for each lysine residue) were used, in addition to variable modifications of methionine oxidation (+16Da) and acetyl serine (+42Da; allows for the occasional acetylation at serine residues; Figure 2.3).

N-terminal peptide identifications with ion scores above the reported threshold were taken as confident assignments. However, N-terminal peptide identifications with scores under the given threshold were examined individually (manual inspection of MS/MS data), and providing a strong ion series (b or y ion) was observed, these peptides were also allowed.

Initially the N-terminal MS/MS data was used to interrogate the SwissProt database to obtain annotated protein identifications. However, N-terminal peptide identifications made using the SwissProt database are restricted to proteins that have undergone no cleavage at the N-terminus. In order to identify N-termini that have undergone SP removal the same MS/MS data was searched using Mascot against specialised N-terminal databases, constructed in house (see Section 2.11.4), for the species studied.

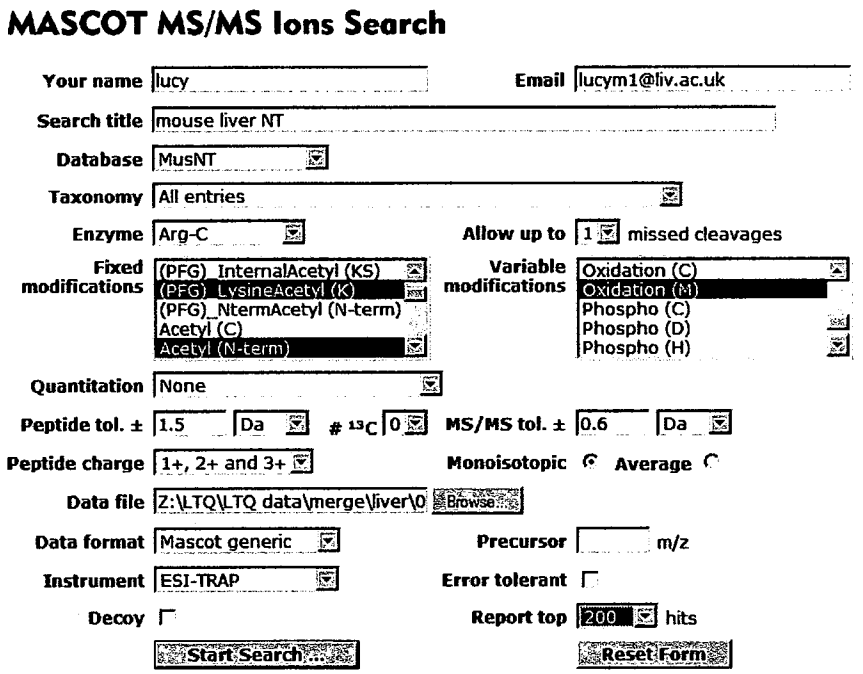


Figure 2.3. MASCOT MS/MS search form.
The MS/MS ions search accepts data in the form of peak lists containing mass and intensity pairs, in this case, the data is in the form of a Mascot generic file (mgf). The parameters shown here are tailored for a complex N-terminal preparation.

Screening for non N-terminal peptides

To screen for non N-terminal (internal and C-terminal) peptides, the N-terminal datasets were searched using the same parameters as when searching N-terminal peptides. However, as an alternative of using N-acetylation as a fixed parameter, this modification was set to variable to allow for non-acetylated, internal or C-terminal, Arg-C peptides. The non N-terminal matches were examined using the same criteria as for the N-terminal peptides.

2.11.4 Construction of N-terminal databases

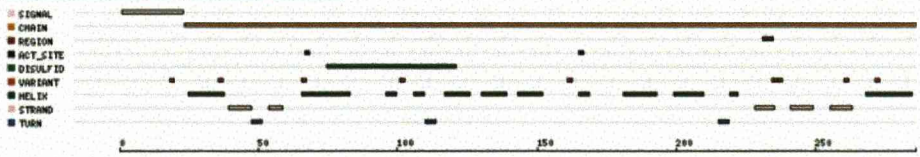
To create various N-terminal databases all *E. coli, S. cerevisiae, Mus musculus* and *Homo sapiens,* sequences were extracted from SwissProt as a FASTA file. The databases were pre-processed (using software written in house by Robert Beynon) to create Arg-C N-terminal peptides, reading where possible feature entries (FT lines), to remove signal peptides and to allow for additional processing events such as signal peptide removal (for an example SwissProt entry see Figure 2.4). Additionally, if the N-terminal amino acid was a methionine residue, two entries were created, one in which the methionine residue was included in the entry and a second in which the N-terminal methionine was removed. This database was then searched using our in-house Mascot server.

2.12 Modeling Expected Isotope Distribution Profiles of MIDAR Derivatives

To model the expected isotope distribution patterns, isotope envelope profiles were calculated for peptides at 1000Da, 2000Da and 3000Da (the masses typically encountered in a proteomics experiment). For each peptide, the isotope profile was calculated using the "averagine" atomic composition for a typical peptide (Senko *et al.,* 1995), using the tool "MS-Isotope" that is part of the Protein Prospector Suite (http://prospector.ucsf.edu). These isotope profiles were used to calculate the expected mass spectrum for the three peptides, with 0, 1, 2 or 3 amino groups and a reagent mixture of 10% [$^{13}C_4$] acetic anhydride and 90% [$^{2}H_6$] acetic anhydride (when the acetyl group is incorporated, the mass gains are 44Da and 45Da for the two labelled variants, respectively).

Features

Features compressed | Features expanded



Key	Begin	End	Length	Description	RESID
SIGNAL	1	23	23		
CHAIN	24	286	263	Beta-lactamase TEM. /FTId=PRO_0000016978.	
REGION	232	234	3	Substrate binding.	
ACT_SITE	68	68	1	Acyl-ester intermediate.	
ACT_SITE	166	166	1	Proton acceptor.	
DISULFID	75	121	47		RESID:AA0025
VARIANT	19	19	1	L -> F (in TEM-4).	
VARIANT	37	37	1	Q -> K (in TEM-2, TEM-3, TEM-8, TEM-16 and TEM-24).	
VARIANT	67	67	1	M -> L (in IRT-4).	
VARIANT	102	102	1	E -> K (in TEM-3, TEM-4, TEM-6, TEM-8, TEM-16 and TEM-24).	
VARIANT	162	162	1	R -> H (in TEM-6 and TEM-16).	
VARIANT	162	162	1	R -> S (in TEM-5, TEM-8 and TEM-24).	
VARIANT	235	235	1	A -> T (in TEM-5 and TEM-24).	
VARIANT	236	236	1	G -> S (in TEM-3, TEM-4 and TEM-8).	
VARIANT	237	237	1	E -> K (in TEM-5 and TEM-24).	
VARIANT	261	261	1	T -> M (in TEM-4).	
VARIANT	272	272	1	N -> D (in IRT-4).	
HELIX	25	38	14		
STRAND	40	48	9		
TURN	49	51	3		
STRAND	54	59	6		
HELIX	67	69	3		
HELIX	70	83	14		
HELIX	97	99	3		
HELIX	107	109	3		
TURN	111	113	3		
HELIX	117	126	10		
HELIX	130	139	10		
HELIX	143	152	10		
HELIX	166	168	3		
HELIX	181	193	13		
HELIX	199	210	12		
TURN	216	218	3		
HELIX	219	222	4		
STRAND	228	235	8		
STRAND	241	249	9		
STRAND	255	263	9		
HELIX	268	284	17		

Figure 2.4. SwissProt feature table for *E. coli* β-lactamase (P62593).
The FT (Feature Table) lines provide a precise but simple means for the annotation of the sequence data. The table describes regions or sites of interest in the sequence. In general, the feature table lists signal peptide cleavage sites, posttranslational modifications, binding sites, local secondary structure or other characteristics reported in the cited references. Sequence conflicts between references are also included in the feature table.

3. DEVELOPMENT OF NOVEL STRATEGIES FOR N-TERMINAL PEPTIDE ISOLATION

3.1 INTRODUCTION

Chemical modification of biological constituents has been used for many years in protein chemistry (reviewed in Wong, 1991). Proteins can be covalently modified in a variety of ways to suit the purpose of a particular study or goal. The most common targets for protein labelling are primary amines, which occur mainly on lysine side chains (ε-amino groups) and on the N-terminus (α-amino group) of protein molecules. Derivatisation of primary amines in proteins and peptides provides the basis for many widely used proteomic strategies.

3.2 EDMAN DEGRADATION

All current enzymic ladder sequencing methods are based on the strategy developed by Pehr Edman in 1949, referred to as "Edman degradation" (Edman, 1949). This strategy utilises the isothiocyanate derivative phenylisothiocyanate (PITC; Edman's reagent) for the stepwise degradation of peptides. In this process the derivatised amino acid is removed from the protein without modifying the remaining peptide chain, thus allowing sequential degradation of the peptide. Each derivatised amino acid (phenylthiohydantoin [PTH] amino acid) is identified by one cycle of Edman chemistry followed by HPLC to analyse the PTH amino acid. This method led to the development of an automated sequencing instrument in 1967 (Edman and Begg, 1967), and shortly after to the construction of the solid-phase sequencer, in which the degraded peptides were covalently attached to a solid support (Laursen, 1971).

The Edman reaction is divided into three stages: (1) coupling of reagent to the amino terminal; (2) cleavage of the modified amino acid and (3) conversion to the PTH amino acid (Figure 3.1). In the coupling reaction, phenylisothiocyanate (PITC) reacts with the amino acid residue at the N-terminus under basic conditions to form a phenylthiocarbamyl derivative (PTC-protein). At pH9 coupling is favoured at α-amino groups and occurs within 15-30min. In the cleavage reaction, TFA is used to cleave the first amino acid to yield an anilinothiozolinone derivative (ATZ-amino acid) and the n-1 polypeptide. The shortened n -1 polypeptide has a new N-terminal amino acid with a reactive amino group, which can undergo another round of coupling and cleavage. In the conversion stage, the unstable ATZ amino acid is converted to a

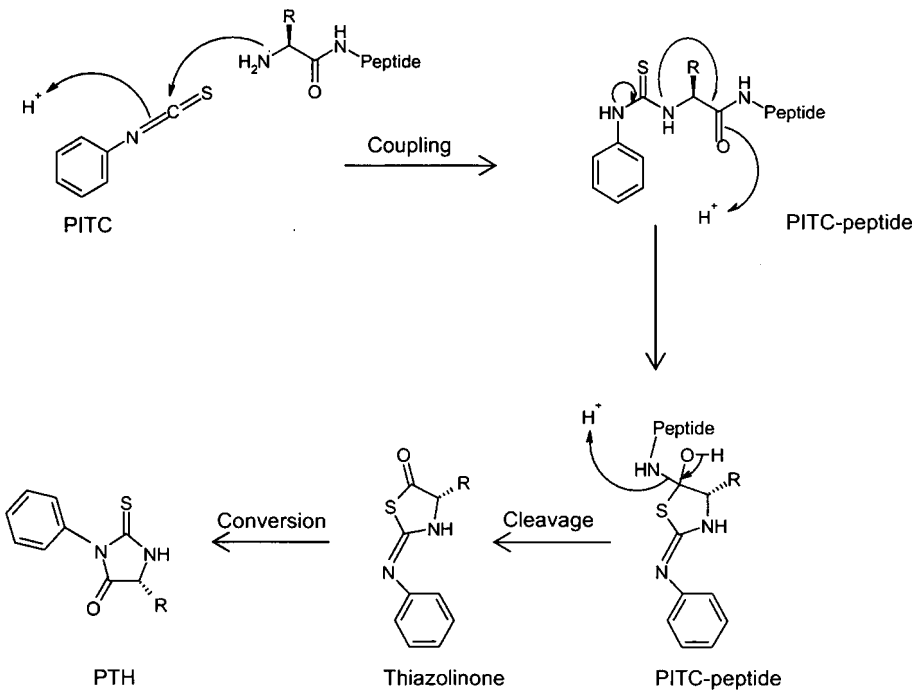


Figure 3.1. The Edman reaction.
The first reaction (coupling) modifies the amino terminus by the addition of phenylisothiocyanate (PITC) to the amino group. The resulting phenylthiocarbamyl (PTC) protein is then treated with an anhydrous acid in a second reaction (cleavage) which allows the sulfur from the PTC group to react with the first carbonyl carbon in the protein chain. This cyclisation reaction results in the removal of the first amino acid as an anilinothiozolinone (ATZ) derivative and leaves the next amino acid in the protein exposed for the next round of PITC coupling. In a third reaction (conversion), the ATZ amino acid is converted to a phenylthiohydantoin (PTH) amino acid in aqueous acid. The PTH is more stable than the ATZ and can be easily analysed. This process may be continued until the limitations of the chemistry or the sample preclude further analysis.

stable phenylthiohydantoin derivative (PTH-amino acid) with anhydrous acid. The PTH-amino acid is transferred to a RP C18 column for detection at 270nm. A standard mixture of 19 PTH-amino acids is also injected onto the column to provide standard retention times of the PTH amino acids.

There are a number of limitations associated with Edman chemistry, including low sensitivity (detection limit of 1pmol; Shively, 2000) and the requirement for purified proteins. However, the major limitation is inhibition of the reaction by blocked N-terminal peptides. The majority of eukaryotic proteins are blocked by N^{α}-acetylation and although it is sometimes possible to unblock N^{α}-acetylated proteins using acylaminoacyl-peptide hydrolase (Krishna *et al.* 1991; Farries *et al.*, 1991) this remains a major problem for Edman chemistry.

Despite its associated limitations, Edman sequencing has proved a valuable tool in proteomic research, in defining the true N-terminal regions of proteins. In a pioneering study, Edman sequencing was used in combination with 2-D SDS-PAGE to identify the true N-terminal protein sequences of 223 unique proteins from *E. coli* (Link *et al.*, 1997). In this study 39% of the identified proteins were found to have undergone NME and 24% SP removal. In accordance with published data, the initiator methionine was cleaved when the penultimate residue was alanine or serine. However, when the penultimate amino acid was threonine, glycine or proline, cleavage was variable. Valine did not permit cleavage.

3.3 Amine reactive isobaric tagging reagents

A recently developed NHS ester-based strategy for isobaric, stable isotope labelling of peptides (iTRAQ; Ross *et al.*, 2004) is now an established technique in proteomics (Zieske 2006; Aggarwal *et al.*, 2006). This strategy was originally developed using a multiplex set (4-plex) of isobaric reagents that modify exposed amino groups (α and ε) on the target proteins. More recent versions of the iTRAQ strategy use a greater number of isobaric reagents (8-plex) (Choe *et al.*, 2007). In both cases, the derivatised peptides are indistinguishable in MS, but following CID yield signature or reporter ions that can be utilised to identify and quantify individual members of the multiplexed set. The reagents are non-polymeric isobaric tagging agents consisting of a reporter group, a balance group and a peptide reactive group (Figure 3.2a). The reporter group differs in mass in each of the variants from m/z 114.1 to 117.1 while the balance group ranges in mass from 28 to 31 Da, thus yielding a constant combined mass of 145.1Da for each of the four reagents.

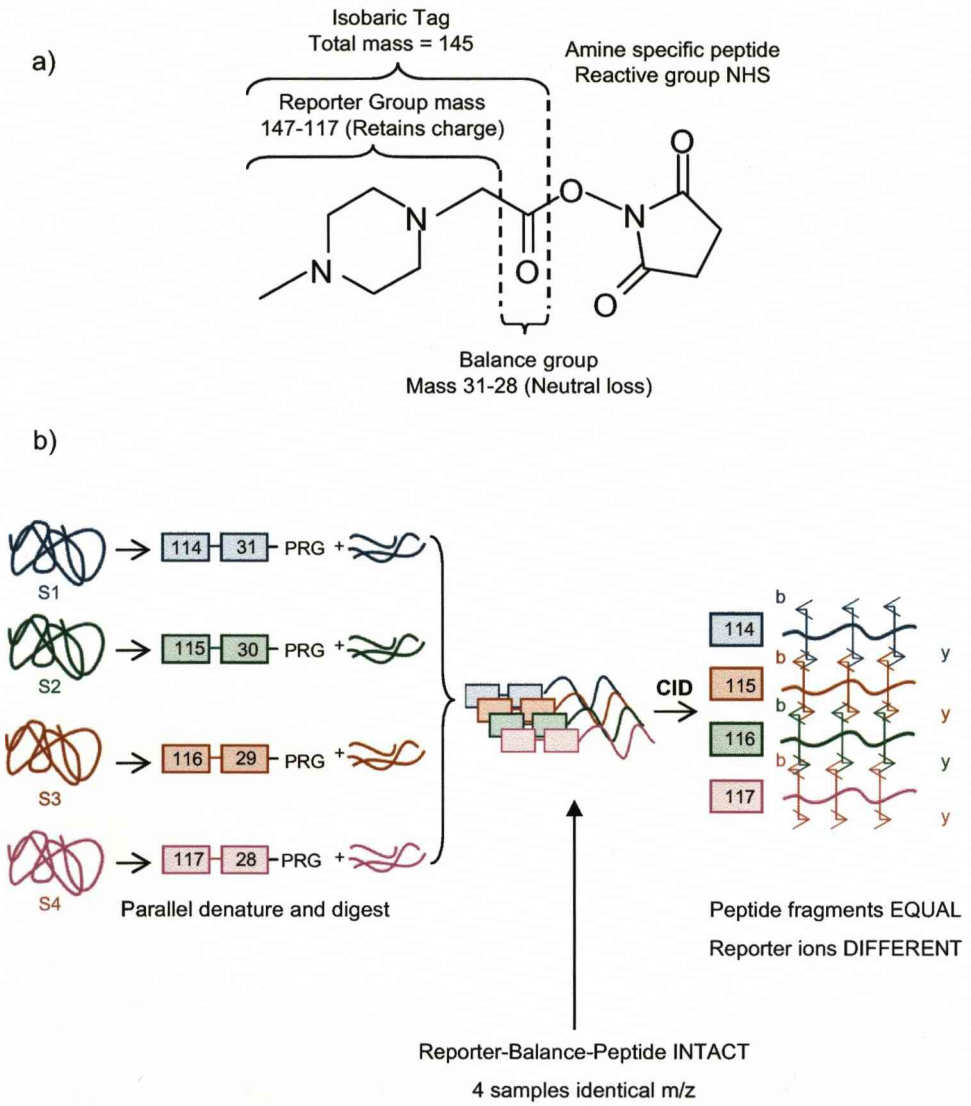


Figure 3.2. Components of the multiplexed isobaric tagging strategy.
The complete molecule consists of a reporter group (based on N-methylpiperazine), a mass balance group (carbonyl) and an amine reactive group (NHS ester; a). In the multiplexed reaction of four different samples (S1-S4; b), the reporter group ranges in mass from m/z 114.1 to 117.1, while the balance group ranges in mass from 28 to 31Da. As the combined mass remains constant (145.1Da), there is no difference observed between samples at the MS stage of analysis. Following fragmentation by CID, the four reporter group ions appear as distinct masses (114 -117Da), where as all other fragment ions (for example, b and y ion series) stay the same. The relative concentration of the peptides is thus deduced from the relative intensities of the reporter ions.

In a typical iTRAQ experiment, four different biological samples are proteolysed and the resulting peptides are derivatised with a distinct, differentially labelled tag. The labelled peptides from each sample are mixed and analysed by LC-MS and MS/MS. The isobaric nature of these reagents means that the same peptide in each sample appears as a single peak in the MS spectrum which reduces complexity of the spectra compared with other techniques such as ICAT (Shiio and Aebersold, 2006). Following CID the multiplexed isobaric tags produce abundant signature ions at m/z 114.1, 115.1, 116.1 and 117.1, the relative intensities of these ions correspond to the proportions of the labelled peptides (Figure 3.2b).

The iTRAQ approach has been successfully applied to a variety of biological samples in order to simultaneously identify and quantify, in relative terms, proteins. Quantitative studies have been performed on low abundance proteins and transcription factors in *E. coli* (Aggarwal *et al.*, 2005), human saliva (Hardt *et al.*, 2005) and human fibroblasts (Cong *et al.*, 2006).

3.4 Amino group derivatisation

When performing coupling reactions of amino groups with amine reactive agents, it is important to create a reaction environment that will favor adequate reactivity. Many amine reactive reagents, for example acid anhydrides, are unstable in aqueous solutions due to hydrolysis and depletion of the reagent in these conditions is unavoidable. However, the rate of hydrolysis is dependent on the particular reagent, temperature, pH and buffer composition.

Derivatisation reactions are most commonly performed in phosphate, bicarbonate/ carbonate, or borate buffers at concentrations between 50-200mM. Other buffers may also be used if they do not contain primary amines. N-2-hydroxyethylpiperazine-N'-2-ethanesulfonic acid (HEPES), for example, can be used because it contains only tertiary amines. Tris is a primary amine, which makes it an unacceptable buffer for use with an amine reactive agent. However, a large excess of Tris, at neutral-to-basic pH may be added at the end of a reaction to quench residual reagent. Glycine also contains a primary amine and may be used in a similar manner. Reactions are typically performed at pH7-9, between 4°C and room temperature and from 30min to 2h. Reagents are typically used at 2-50 fold molar excess depending on the concentration of the protein to be modified.

The risk factors associated with the stability of amine reactive agents along with strategies for minimising their effects are listed in Table 3.1.

Problem	Solution
Instability of reagent	Prepare all reagents immediately prior to derivatisation
Acidification of reaction environment by hydrolysis of excess reagent	Use reaction buffer able to maintain optimal pH for the duration of the reaction
Competition from other reagents in the reaction mixture	Ensure that no amine-containing buffers, for example Tris, have been added to the reaction mixture

Table 3.1. Risk factors commonly associated with amine modification *in vitro*.

3.4.1 Acylating agents

The most commonly used amine specific reagents for protein modification are acylating agents. These agents are compounds that contain an activated acyl group in which the nucleophile on the primary amino group attacks at the carbonyl carbon, displacing a leaving group. Commonly used acylation reagents include: acid anhydrides, isocyanates, isothiocyanates, N-hydroxysuccinimidyl (NHS) and other activated esters (Figure 3.3). All nucleophiles contained within the protein molecule are susceptible to acylation. Under aqueous conditions, O-acylation of hydroxyl-containing amino acid residues (serine, threonine, and tyrosine) may be a significant side reaction. An important distinction between O-acylation and N-acylation is that O-acylation is easily reversible. Hydroxylamine (or strongly alkaline conditions) is sufficient to remove acyl groups from hydroxyl-containing amino acid residues (Miller, 1996).

An acid anhydride is an organic compound that has two acyl groups bound to the same oxygen atom (Moss *et al.,* 1995). When the two acyl groups are directly derived from a carboxylic acid (the most common case), the general formula is RC(O)OC(O)R. Symmetrical acid anhydrides of this type are named by replacing the word *acid* in the name of the parent carboxylic acid by the word *anhydride*.

Acetic anhydride is an organic solvent with the formula $(CH_3CO)_2O$ (Figure 3.4a). It is the most commonly used acetylation reagent in organic synthesis and it is widely used in the production of drugs such as heroin (diacetylmorphine; Klemenc, 2002). Aqueous solutions of acetic anhydride have limited stability because, like most acid anhydrides, acetic anhydride hydrolyses to give acetic acid (Figure 3.4b).The instability of acetic anhydride in aqueous conditions will cause the pH of a reaction to fall significantly. Therefore, when using acetic anhydride as a derivatising agent in a pH sensitive environment, it is necessary to use a system with adequate buffering capacity.

Both isocyanates and isothiocyanates are extremely reactive. Reactions mainly occur through addition to the double bond joining the carbon and nitrogen atoms (Figure 3.3). The nucleophile (protonated amino group) attacks the electrophilic carbon atom and the active hydrogen is added to the nitrogen atom. Isocyanates have the reactive group -N=C=O, whereas isothiocyanates have the chemical group -N=C=S, formed by substituting sulfur for oxygen in the isocyanate group. As discussed earlier, phenylisothiocyanate is a derivative of isothiocyanate that has important implications in Edman sequencing.

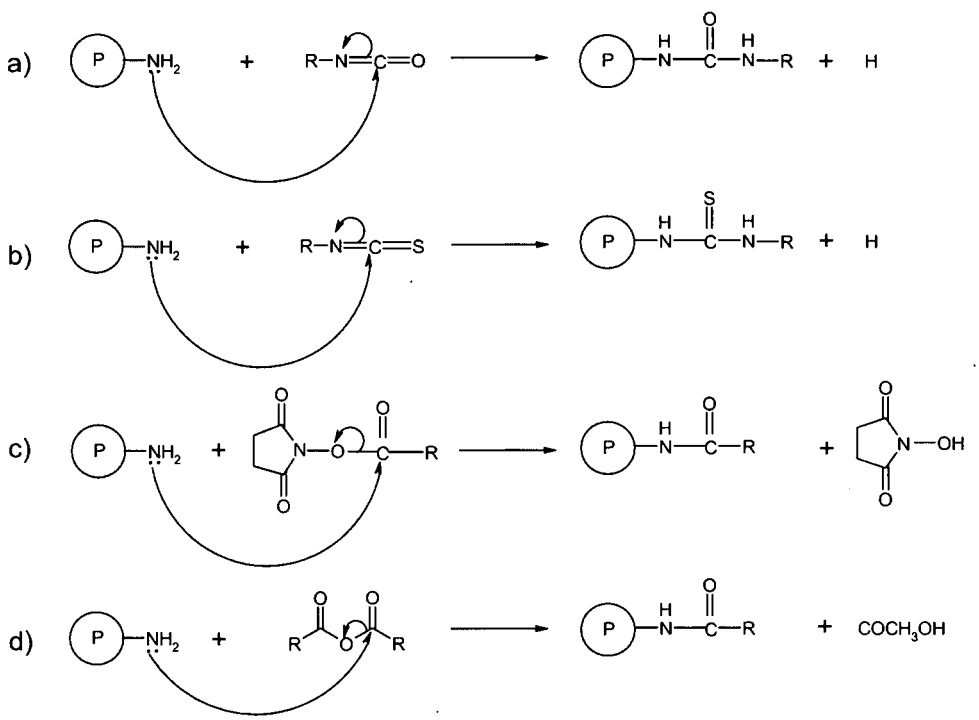


Figure 3.3. Reactions of commonly used acylation reagents with α-amino groups. Reaction with: (a) Isocyanate; (b) isothiocyanate; (c) NHS ester and (d) acid anhydride.

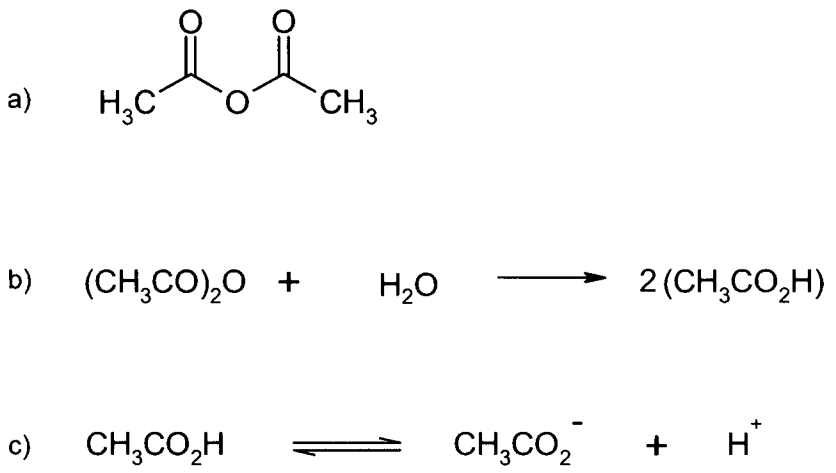


Figure 3.4. Structure and hydrolysis of acetic anhydride.
a) Acetic anhydride consists of two acetyl groups bound to the same oxygen atom.
b) Aqueous solutions of acetic anhydrides hydrolyse to produce acetic acid.
c) Acetic acid dissociates to release [H^+], therefore adequate buffering is important.

3.4.2 Biotin NHS esters

Biotin, also known as vitamin H or vitamin B7, is a relatively small molecule (226Da) that can be covalently attached to selected amino acid residues, in most cases, without a significant effect on biological activity. Derivatisation of proteins and peptides through biotinylation has increased significantly over the past two decades (Wilchek and Bayer, 1988; Wilchek and Bayer, 1989). The addition of biotin groups can be exploited using the extremely high affinity of biotin for the tetrametic proteins avidin, which is a glycoprotein found in egg white and also streptavidin from the microorganism *Streptomyces avidinii*, with a dissociation constant in the order of 10^{15}mol/l (Wilchek and Bayer, 1988). Biotin can be covalently linked to proteins or peptides using one of the many commercially available biotinylation reagents, the most popular is biotinyl NHS-ester (NHS-biotin). Figure 3.5 shows the structure of NHS-biotin and the mechanism of reaction with primary amines. NHS ester-mediated derivatisation of primary amino groups was first described by *Becker et al.*, in the study of biotin transport in yeast (Becker and Wilchek, 1972; Becker *et al.*, 1971). Since then many other applications have been developed, for example the previously described techniques of ICAT and iTRAQ.

3.5 N-TERMINAL PEPTIDE ISOLATION

Amino group derivatisation is also routinely applied to N-terminal peptide purification strategies (reviewed in Chapter 1). In contrast to Edman degradation, the more recent chemistries for N-terminal isolation are not impeded by N^{α}-acetylation. In addition to the benefits gained by simplification, these strategies have a useful application in determining the true nature of protein N-termini.

This chapter describes the development of two novel 'positional proteomic' approaches for enrichment of N-terminal peptides from complex protein mixtures using different types of reagents in the post-proteolytic stages. Both approaches involve an initial acetylation step, which is necessary to block the α-amino group on the true N-terminal of the proteins and also to block and protect the ε-amino group on the side chains of lysine residues. The success of this initial acetylation step requires the optimisation of reaction conditions to maximise the yield of derivatised proteins. Before developing the N-terminal isolation procedure, it is necessary to carry out a series of experiments designed to optimise the reaction conditions for acetylation. The initial experiments will involve monitoring the change in pH of sodium carbonate buffer on addition of the acetylation reagents using pH paper. Once pH is optimised the next stage will be to modify a model peptide (ACTH 1-17) with two

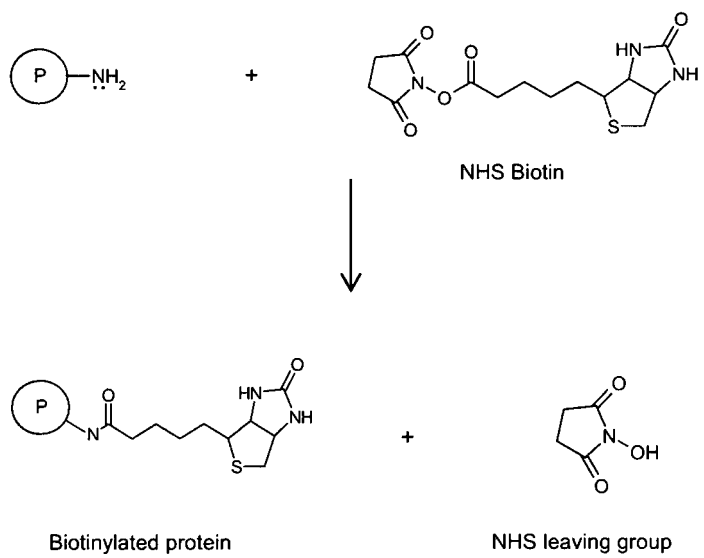


Figure 3.5. Modification of protein amino groups with NHS biotin.
The nucleophile (amino group) is attacked by the electrophile (carbonyl carbon) from the NHS group.

reagents: acetic anhydride and sulfo-NHS acetate. Both of these reagents introduce an acetyl group to exposed amino groups on protein molecules, however, the chemistries in which they achieve this modification differ (Figure 3.3). The two reagents will be optimised using the modification of ACTH 1-17 in a time course experiment. Once the conditions for acetylation are established the reaction will be applied to a purified protein. The purpose of this experiment is to determine if acetylation of an intact protein in its native state is feasible. If successful then acetylation can be applied to real biological samples in their native state, followed by further stages of the N-terminal isolation procedure.

Following acetylation and proteolysis, the first approach for N-terminal isolation involves biotinylation of the peptide mixture using NHS-biotin. All internal peptides become biotinylated, however, the acetylated N-terminal peptides are resistant to modification, providing a basis for segregation. Removal of the biotinylated internal peptides, through interaction with streptavidin, results in a preparation enriched in N-terminal peptides. The unbound material containing the N-terminal peptides can be directly analysed by MS (Figure 1.17). As each protein yields a single signature peptide from each protein, the resultant mixture, in theory, should have the same level of complexity as the initial unproteolysed proteome (McDonald *et al.*, 2005).

To summarise approach one: samples are dialysed into a compatible (non-amine containing) buffer prior to acetylation. In some cases the protein mixture can be generated using a compatible buffer which removes the need for dialysis. Excess acetylation reagent is removed using amine scavenger beads (structure represented in Figure 3.6), which facilitates the removal of excess reagent without adding amines to the mixture and can be separated from the sample by centrifugation. The proteins are then concentrated by acid precipitation and the pellet washed prior to proteolysis. The proteolysed mixture is biotinylated using NHS-biotin which adds a biotin moiety to the newly formed α-amino groups on the N-termini of the internal peptides. The peptide mixture, which now consists of biotinylated internal peptides and acetylated N-terminal peptides, is passed over streptavidin Sepharose, which binds to and retains the internal peptides. The derivatisation reactions used in this approach are initially optimised using model peptides and purified proteins, before being applied to the soluble proteins of mouse skeletal muscle. This protocol, although effective, requires multiple rounds of peptide purification following each derivatisation stage to separate the peptides from the excess reagents used which, in turn, results in a reduced yield of material.

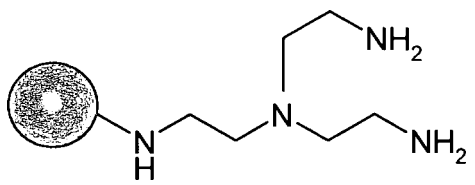


Figure 3.6. Structure of Tris(2-aminoethyl)amine, polymer-bound.

The second approach consists of enhanced methodologies to minimise the number of processing steps, yet maximise yield. The most significant enhancement to the method is the elimination of the biotinylation step. All pre-proteolytic steps remain the same as in the first approach, including the initial acetylation stage to block α-amino groups on the intact proteins. However, following proteolysis, instead of targeting internal peptides by biotinylation and removing them with streptavidin, a commercially available amine reactive immobilised reagent (NHS-activated Sepharose) is used to react with and retain internal peptides in one step. NHS-activated Sepharose is efficient, both in respect of amine binding and subsequent leakage of bound amines (Van Sommeren *et al.*, 1993). The acetylated peptide mixture is incubated with the NHS-Sepharose by centrifugation and analysed without further treatment (Figure 3.7; McDonald and Beynon, 2006).

To clarify, all steps in approach two, up to and including proteolysis are the same as in the first approach described above. The proteolysed mixture is exposed to NHS-activated Sepharose, which binds to and retains all peptides with unblocked α-amino groups i.e. all internal peptides. This strategy for N-terminal isolation will initially be performed on the same sample as the first method (mouse skeletal muscle soluble fraction), before being applied to the more complex proteomes of mouse liver, *S. cerevisiae* and *E. coli* cell lysate, to obtain global proteomic datasets from three distinct biological samples. In all cases, samples will initially be subjected to MALDI-ToF MS to assign N-terminal peptides to abundant species in the mixture. To obtain a greater number of identifications, samples will be fractionated by RP-HPLC over an extended three hour gradient, and analysed by MS/MS on the LTQ ion trap instrument.

The strategies for N-terminal purification described in this chapter share similarities with other methods available in the literature (see Table 1.2). However, our strategy represents the simplest set of chemistries for identifying both forms of N-terminal peptides (N^{α}-acetylated and unblocked) in complex eukaryotic and prokaryotic proteomes. In contrast to the alternative methods outlined in Chapter 1 (Section 1.9.2), our methods do not contain the initial reduction and alkylation stages required to reduce and block disulphide bonds. Therefore, the two strategies demonstrated in this chapter are only suitable for the analysis of soluble protein fractions.

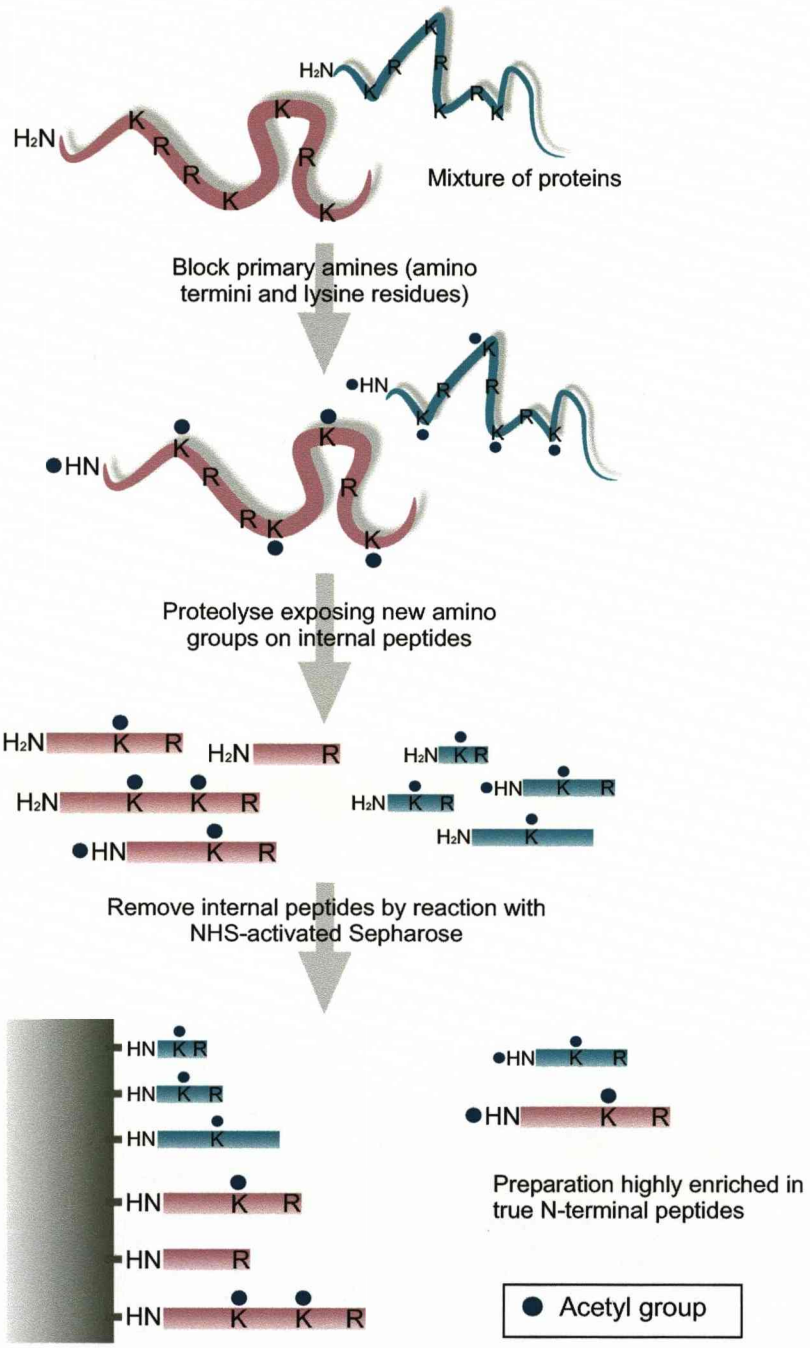


Figure 3.7. Scheme outlining the chemistry involved in N-terminal purification using the NHS-Sepharose method.
Free amino groups (α and ε) are acetylated prior to proteolysis, which results in a mixture of N-terminally acetylated and non-acetylated (internal) peptides. Subsequent incubation of the peptide mixture with NHS-Sepharose results in a preparation enriched in N-terminal peptides.

3.6 Determination of N-terminal acetylation state

Although rare in prokaryotes, amino terminal modification by acetylation is extremely common in eukaryotes. Around 50% of the proteins in fungi are N^{α}-acetylated (Lee *et al.,* 1989b) and as many as 90% in animals (Polevoda and Sherman, 2000).

Unlike NME, the rules governing N^{α}-acetylation are not clear-cut. Eukaryotic proteins which are susceptible to N^{α}-acetylation have a variety of different N-terminal sequences with no simple consensus motif and with no dependence on a single type of amino acid residue (Polevoda *et al.*, 1999). In accordance with NME, N^{α}-acetylation is dependent on the nature of the second encoded residue. Both alanine and serine are the most commonly N^{α}-acetylated amino acids, in addition to the uncleaved methionine residue when the penultimate residue is either aspartic or glutamic acid (Flinta *et al.,* 1986).

Using the conventional N-terminal purification methods described here it would not be possible to distinguish between chemically derived and naturally occurring acetyl groups on N-termini, as both acetyl groups are identical and can not be distinguished by mass. Therefore, in order to discriminate between naturally and chemically acetylated N-termini, the positional proteomics protocol will be adapted using a deuterated from of acetic anhydride ($(C[^2H_3]CO)_2O$). Under these circumstances a chemically acetylated (*in vitro* acetylated) N-terminal will have a mass difference of 45Da and a naturally acetylated (*in vivo* acetylated) N-terminal will have a mass difference of 42Da. When analysed by MS, the difference in mass will be exploited to determine the acetylation status of the true N-terminal peptide. When applied to the set of complex proteomes (mouse liver, *S. cerevisiae* and *E. coli*), this modified version of the positional proteomics strategy will provide a useful survey of the true N^{α}-acetylation status of a large number of mature proteins.

MALDI-ToF MS has been used to determine the N^{α}-acetylation status of 68 ribosomal proteins from a normal strain of *S. cerevisiae* and from three mutant stains each lacking a catalytic subunit of three different acetyltransferases (Arnold *et al.,* 1999). A total of 30 of the 68 ribosomal proteins were N^{α}-acetylated, and 24 of these (80%) had serine at the N-terminus (NatA substrates). In contrast to this time consuming and complex approach to study N^{α}-acetylation, the N-terminal positional proteomic strategy demonstrated in this chapter should be capable of generating the same dataset in a rapid, high-throughput manner.

3.7 Limitations to N-terminal analysis

Protein N-termini are, in general, poorly identified by routine MS analysis (reviewed in Meinnel and Giglione, 2008). Truncated and immature cDNAs or missed 5' introns in genomic DNA frequently result in incorrect N-terminal assignment. In addition, proteolytic events such as N-terminal trimming and SP removal will cause the true N-terminus of a protein to differ from the database entry for the amino acid sequence. Proteins that have undergone proteolytic events will be represented in the database in their precursor form, making it difficult, if not impossible, to identify the mature N-terminal by sequence similarity searching methods alone. Most MS analysis software now includes dedicated tools to address these issues. In addition to the standard enzyme options, software now includes 'no enzyme' or 'nonspecific (half) cleavage' options for a particular enzyme, allowing truncated proteins, including those with abnormal N-terminal cleavage to be identified (Palagi *et al.*, 2006). However, the general use of non-specific cleavage is compromised due to the substantial increase in the amount of data to be searched, leading to a significant increase in analysis time (Craig, 2003). For instance, the 'no enzyme' search must test all possible subsequences of each protein, rather than just (say) tryptic or Arg-C peptides. For a modest size protein of 250 residues, this is an increase of three orders of magnitude in the number of peptides which must be considered.

A separate issue to consider when performing N-terminal searches using existing software tools is the implicit limitation associated with so called "one hit wonders" (Veenstra *et al.*, 2004). This is the phrase coined to describe a protein identification based on one peptide match alone. From a validation point of view, it is preferred that proteins are identified by at least two MS/MS sequenced peptides. However, in the case of an N-terminal preparation each protein is ideally represented by a single peptide. The added information gain obtained from the knowledge of the peptide location within the parent protein is sufficient, in most cases, to validate the assignment. Although, in the case of a poor quality MS/MS spectrum acquired from a low abundant protein in the sample, the confidence of identification will be poor and most likely insignificant. The fundamental problem is that currently MS search software does not incorporate tools to permit the inclusion of positional information. Subsequently, the added information gain relating to the position of the peptide is lost and will not directly contribute to the confidence of the match.

The use of existing proteomic databases (for example, SwissProt) using standard MS searching software, such as Mascot and Sequest, limits the full potential of the positional proteomics approach. In order to maximise the number of identifications obtained from the N-

terminal datasets, it is necessary to construct specialised (processed) N-terminal databases using protein sequence data from the existing repositories. These 'N-terminal databases' will represent mature proteins sequences by allowing for NME and ensuring the removal of signal peptides using the information on predicted cleavage sites. However, these databases will only be representative of true N-terminal peptides provided the algorithms used for SP removal are accurate and they will not allow for other forms of N-terminal modification such as N-terminal trimming through the action of aminopeptidases.

3.8 Aims and Objectives

In this chapter, the development of two distinct strategies for N-terminal purification will be described. The key step to both of these strategies, protein acetylation, will initially be optimised to ensure complete modification. Optimisation will begin with model peptides before being applied to an intact purified protein, then to a complex protein mixture. Once developed fully, the N-terminal isolation strategy will be applied to real proteomes in order to achieve rational simplification of complex protein mixtures.

In addition to simplification, this analysis will provide a survey of the true nature of N-termini in a range of proteomes. To determine the *in vivo* N^{α}-acetylation status of the N-terminal peptides identified in this study, the complex protein mixtures will be acetylated with a deuterated form of acetic anhydride ($(C[^2H_3]CO)_2O$). Progression of the N-terminal isolation strategy, using this reagent, will allow determination of the true N^{α}-acetylation status of the proteins identified.

Finally, the peptides identified from each biological sample will be collated and analysed to enable a survey, including N-terminal amino acid frequency and N^{α}-acetylation status, of a large number of mature proteins.

3.9 Results and Discussion

3.9.1 Optimisation of buffer conditions for acetylation

Acetic anhydride hydrolyses rapidly in aqueous conditions to produce acetic acid (Figure 3.4). For this reason, it is necessary to monitor the pH change of the analytical system in order to ensure adequate buffering capacity. To acetylate both α and ε-amino groups effectively, the pH must be kept above 7 (to ensure that all amino groups are nucleophiles). It is therefore important that hydrolysis of the reagent does not cause the pH to drop below 7, as this will potentially prevent acetylation.

To optimise the reaction environment, acetylation buffer (Na_2CO_3), which is amine free, was prepared at a range of molar concentrations (0.1, 0.2, 0.3, 0.4 and 0.5M) at pH9. Acetic anhydride (1µl/10µmol) was added to individual samples of buffer (50µl) and the pH change monitored using pH paper. The addition of 1µl acetic anhydride caused the pH of the 0.1-0.4M buffers to drop immediately from pH9 down to pH2. This pH drop is due to the high concentration of protons produced by hydrolysis of acetic anhydride (Figure 3.4). In aqueous conditions 10µmol of acetic anhydride hydrolyses to produce 20µmol of acetic acid, which in turn, will require 20µmol of base to buffer the system. To maintain a constant pH, it was necessary to use 50µl of 0.5M Na_2CO_3 (providing 25µmol of base). However, under these conditions it is expected that O-acetylation of serine and tyrosine residues will be extensive.

Unlike acid anhyrides, NHS esters do not hydrolyse rapidly in aqueous conditions. For this reason the addition of sulfo-NHS acetate to an aqueous environment should not result in a substantial drop in pH. The addition of 1mg sulfo-NHS acetate to 50µl of 20mM Na_2CO_3 at pH9, caused only a small pH drop (pH9 to 7), which should be sufficient for amino group derivatisation (Wong, 1991).

3.9.2 Acetylation of model peptides

To establish the most suitable reaction conditions for amino group acetylation, two separate time course experiments were set up using two different acetylation reagents: acetic anhydride and sulfo-NHS acetate. The model peptide ACTH fragment 1-17 was chosen for this experiment. This standard peptide (2092.08Da) contains three lysine residues, two serine residues and one tyrosine residue (Figure 3.8). In addition to the α-amino group, this makes a total of seven possible acetylation sites (four N-acetylation and three O-acetylation sites).

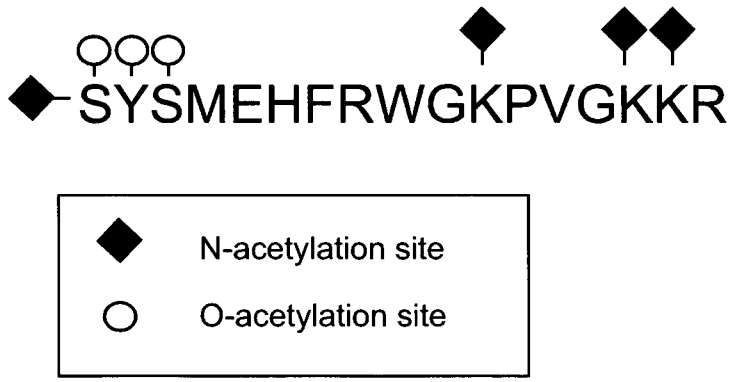


Figure 3.8. Sequence of ACTH fragment 1-17, highlighting potential N and O-acetylation sites.

No. acetylation sites	$[M+H]^+$	$[M+2H]^{2+}$	$[M+3H]^{3+}$
0	2093.09	1047.05	698.37
1	2135.10	1068.05	712.37
2	2177.11	1089.06	726.37
3	2219.12	1110.06	740.38
4	2261.13	1131.07	754.38
5	2303.14	1152.07	768.38
6	2345.15	1173.08	782.38
7	2387.16	1194.08	726.39

Table 3.2. m/z of ACTH fragment 1-17 in multiply acetylated forms.

Table 3.2 indicates the expected m/z values ($[M+H]^+$, $[M+2H]^{2+}$ and $[M+3H]^{3+}$) for the unmodified peptide, all partially acetylated intermediates and the fully modified peptide.

Each time course experiment consisted of 10 samples of ACTH 1-17, each containing 10µg (5nmol) of the peptide. In the first experiment, the peptide was diluted in 0.5M Na_2CO_3, pH9 (50µl) and 1µl acetic anhydride was added to each aliquot. In the second experiment, the peptide was diluted in 20mM Na_2CO_3, pH9 (50µl) and 0.1mg sulfo-NHS acetate was added to each aliquot. The samples were mixed by vortexing and the reactions halted by the addition of 1MTris, pH8 (10µl). Hydroxylamine (1µl/30µmol) was added to the final sample in both experiments to remove potential acetyl groups from serine or threonine residues. Each sample was desalted using a C18 ZipTip prior to MALDI-ToF and ESI-Q-ToF analysis.

Figure 3.9 shows the MALDI-ToF spectra for the acetic anhydride time course experiment. From the masses listed in Table 3.2, it can be seen that immediately after the addition of 1µl acetic anhydride, the peptide was present in three forms, corresponding to the addition of 4, 5 and 6 acetyl groups. After 1min an additional peak appears corresponding to the $[M+H]^+$ value for 7 acetyl groups, indicating that the peptide is acetylated on all possible sites. Following the addition of 1µl (30µmol) hydroxylamine a single peak is present at 2261.13 m/z corresponding to the $[M+H]^+$ value of four acetyl groups. The results were similar for the NHS acetate time course (Figure 3.10). In this case the peptide was not fully N-acetylated immediately after the addition of the reagent; however, acetylation was complete after 30min.

It is not possible to determine the precise location of the acetyl groups from mass alone. In order to establish which amino acid residues have been modified, it is necessary to obtain *de novo* sequence data. Each sample was diluted to 1pmol/µl and infused into the ESI-Q-ToF for MS/MS analysis. Product ion spectra were collected for the unmodified peptide (Figure 3.11), all acetylated intermediates (Figures 3.12-3.14), the fully acetylated peptide (Figure 3.15) and the hydroxylamine treated peptide (Figure 3.16). All ions observed were $[M+3H]^{3+}$. Table 3.2 shows the expected $[M+3H]^{3+}$ values for all modified forms of the peptide. The *de novo* sequence data shows that acetylation occurs initially on the primary amino groups (α-amino group and lysine residues) followed by modification at serine and tyrosine residues (Figure 3.12-3.15). Addition of 1µl hydroxylamine facilitated the removal of acetyl groups from serine and tyrosine, resulting in the presence of a single, fully N-acetylated peptide (Figure 3.16).

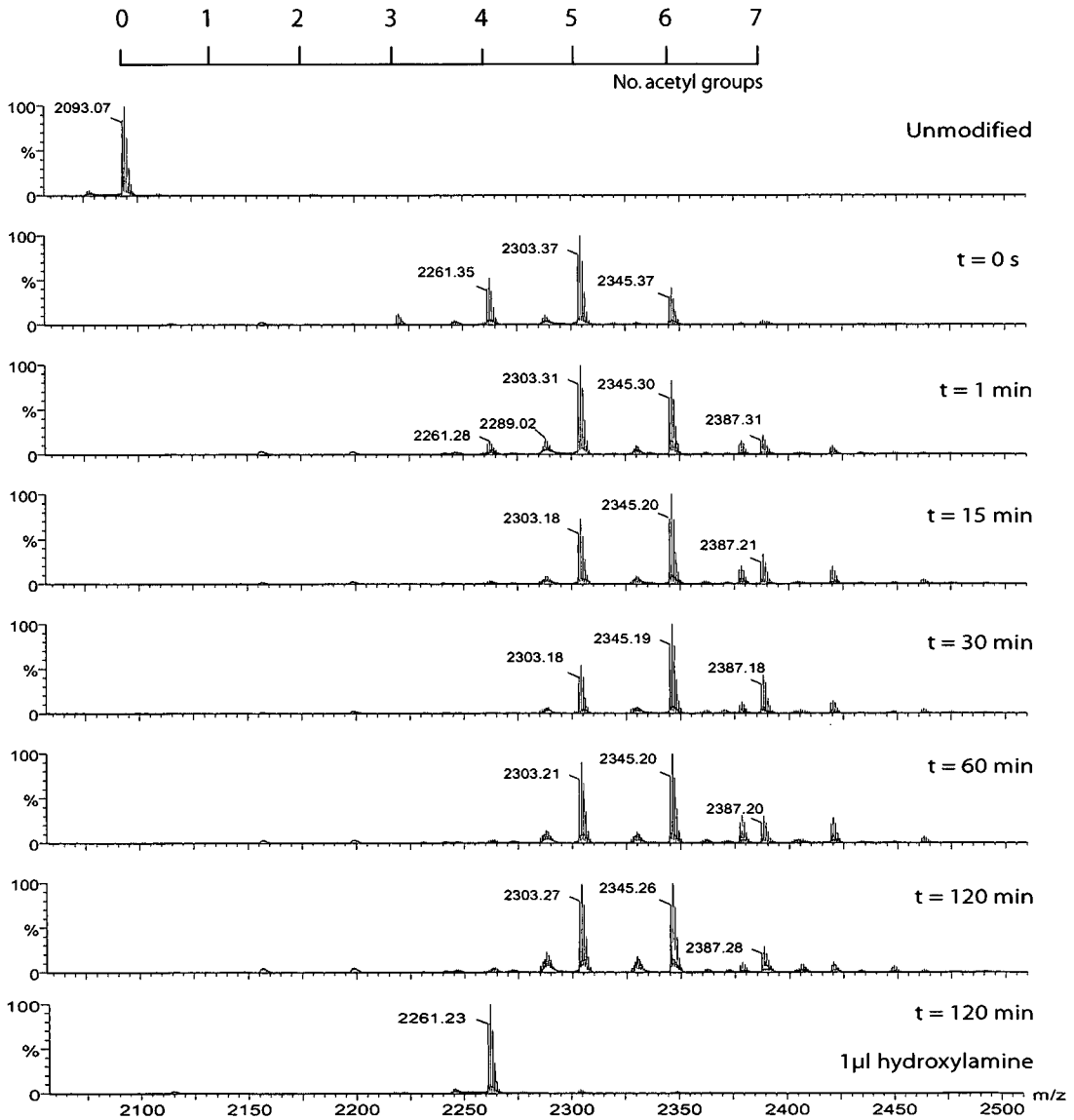


Figure 3.9. ACTH 1-17 acetylation time course (acetic anhydride).

The model peptide ACTH 1-17 (10µg) was acetylated using 1µl acetic anhydride. The reaction was quenched at 0, 1, 15, 30, 60 and 120min by the addition of 5µl Tris (1M, pH8.5). The final sample was treated with 1µl of hydroxylamine to reverse O-acetylation at serine and tyrosine residues. The unmodified and acetylated peptides were desalted on a C18 ZipTip prior to MALDI-ToF MS. $[M+H]^+$ values for the unmodified peptide and acetylated intermediates are represented in Table 3.2.

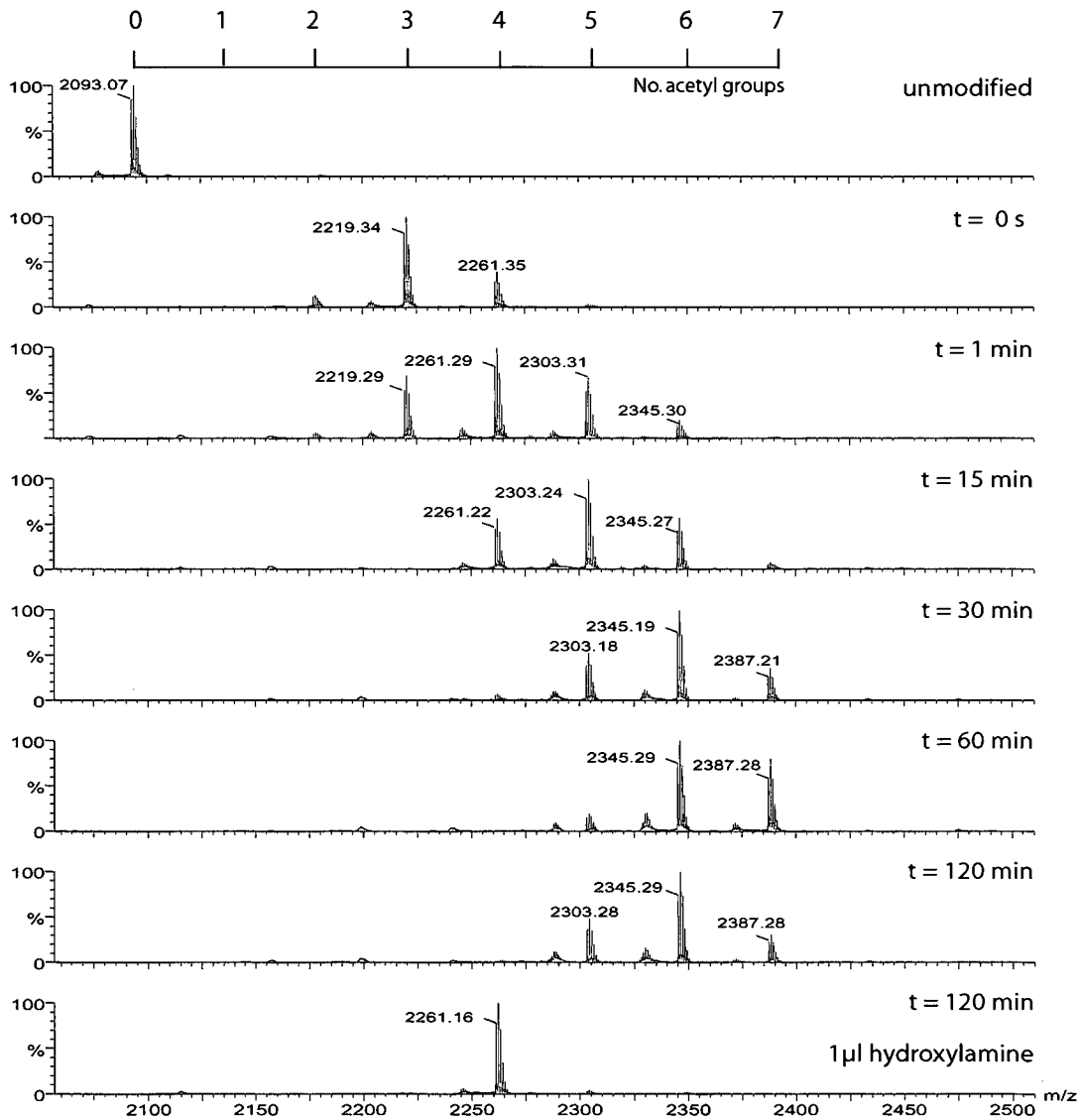


Figure 3.10. ACTH 1-17 acetylation time course (sulfo-NHS acetate).

The model peptide ACTH 1-17 (10µg) was acetylated using 1mg sulfo-NHS acetate. The reaction was quenched at 1, 1, 15, 30, 60 and 120min by the addition of 5µl Tris (1M, pH8.5). The final sample was treated with 1µl of hydroxylamine to reverse O-acetylation at serine and tyrosine residues. The unmodified and acetylated peptides were desalted on a C18 ZipTip prior to MALDI-ToF MS. $[M+H]^+$ values for the unmodified peptide and acetylated intermediates are represented in Table 3.2.

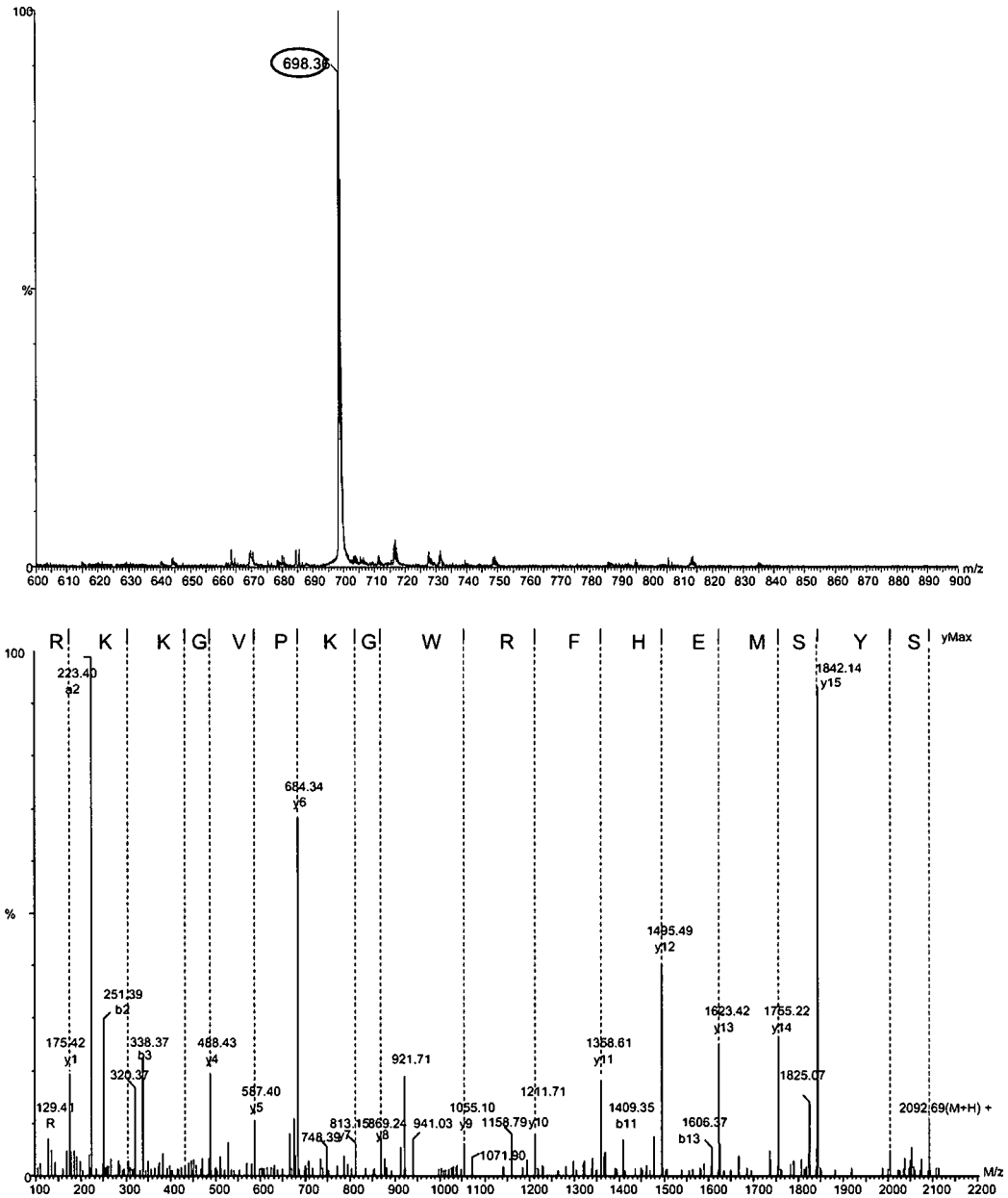


Figure 3.11. Sequence determination of unmodified ACTH 1-17.
The model peptide ACTH 1-17 was diluted in 0.1% (v/v) FA: 50% (v/v) ACN, to give a final concentration of 1pmol/µl. The solution containing the peptide was introduced into the ESI-Q-ToF by direct infusion at a rate of 0.5µl/min. From the MS spectrum (upper panel) the precursor ion at $[M+3H]^{3+}$ 698.36 was selected using the quadrupole fragmented using a collision energy of 30%. The product ion spectra were combined and processed using MaxENT3. The resulting MS/MS spectrum was sequenced *de novo* using the Biolynx software tool (lower panel).

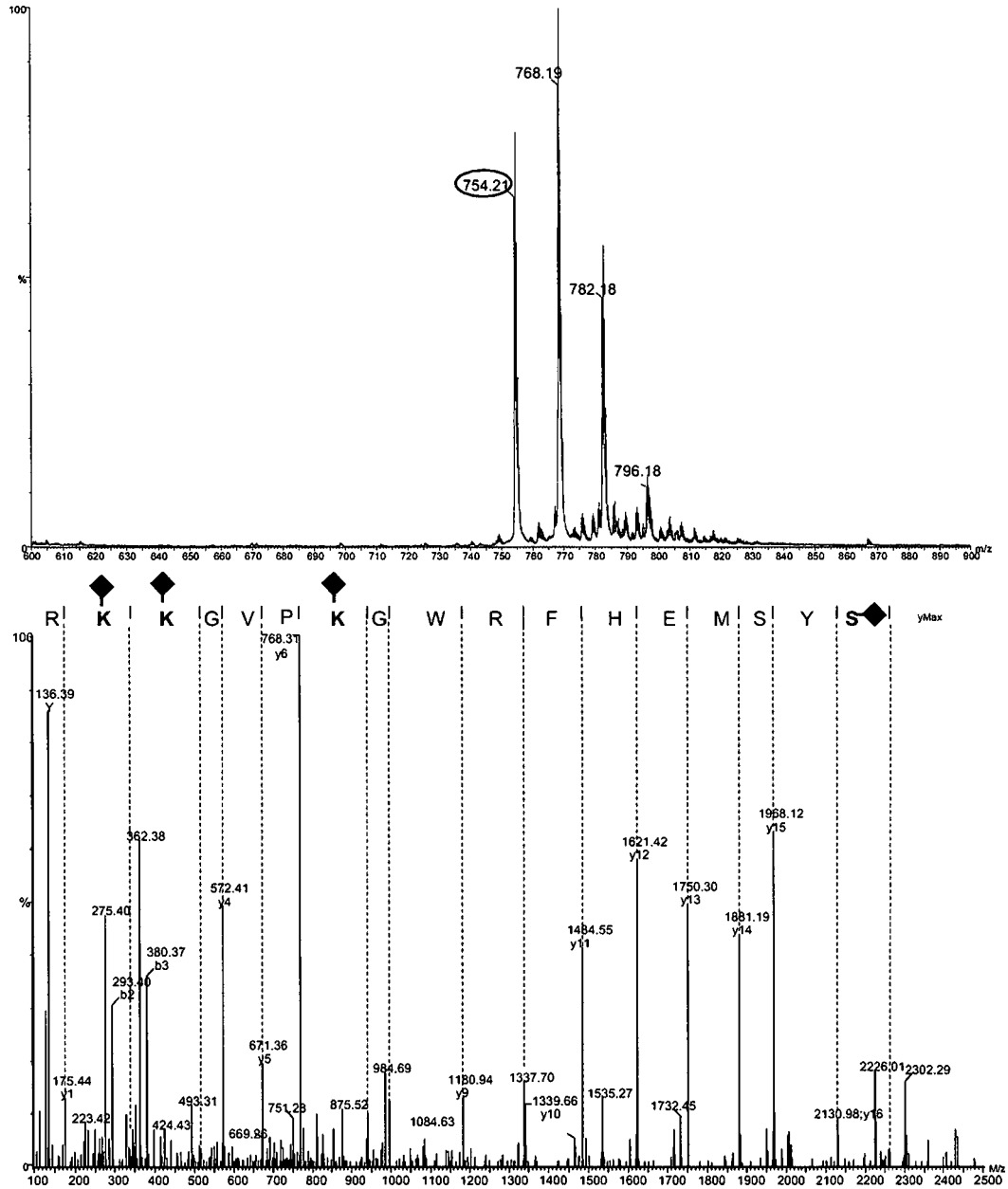


Figure 3.12. Determination of the position of acetyl groups coupled to ACTH 1-17 $[M+3H]^{3+}$ 754.21.

Sample t=0 from the acetic anhydride acetylation time course was desalted using a C18 ZipTip and diluted (1in20) using 0.1% (v/v) FA: 50% (v/v) ACN. The differentially acetylated peptide mixture was introduced into the ESI-Q-ToF by direct infusion at a rate of 0.5µl/min. From the MS spectrum (upper panel), the precursor ion at $[M+3H]^{3+}$ 754.21 was selected using the quadrupole and fragmented using a collision energy of 30%. The product ion spectra were combined and processed using MaxENT3. The resulting MS/MS spectrum was sequenced *de novo* using the Biolynx software tool (lower panel).

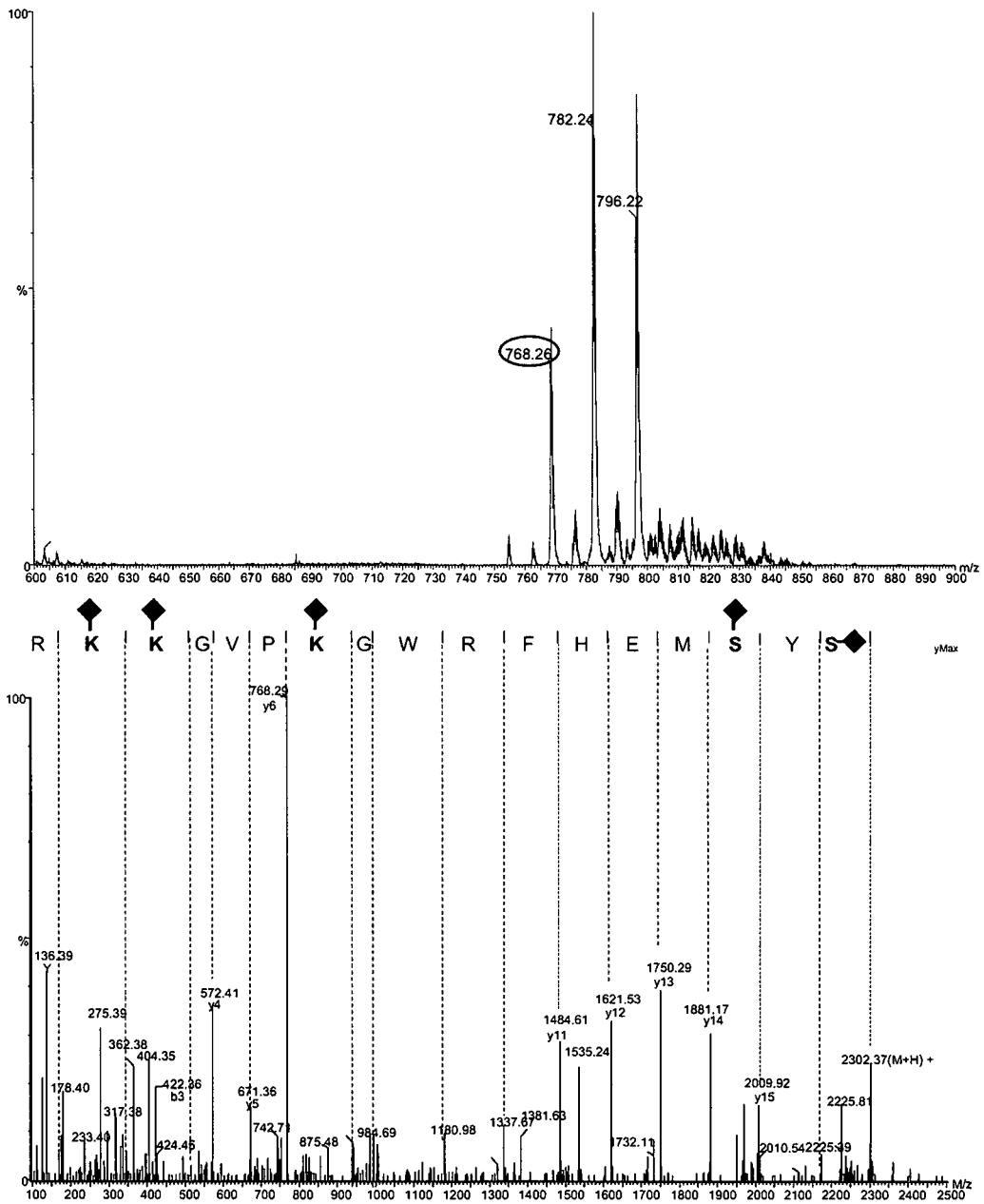


Figure 3.13. Determination of the position of acetyl groups coupled to ACTH 1-17 $[M+3H]^{3+}$ 768.26 .

Sample t=1min from the acetic anhydride acetylation time course was desalted using a C18 ZipTip and diluted (1in20) using 0.1% (v/v) FA: 50% (v/v) ACN. The differentially acetylated peptide mixture was introduced into the ESI-Q-ToF by direct infusion at a rate of 0.5µl/min. From the MS spectrum (upper panel) the precursor ion at $[M+3H]^{3+}$ 768.26 was selected using the quadrupole and fragmented using a collision energy of 30%. The product ion spectra were combined and processed using MaxENT3. The resulting MS/MS spectrum was sequenced *de novo* using the Biolynx software tool (lower panel).

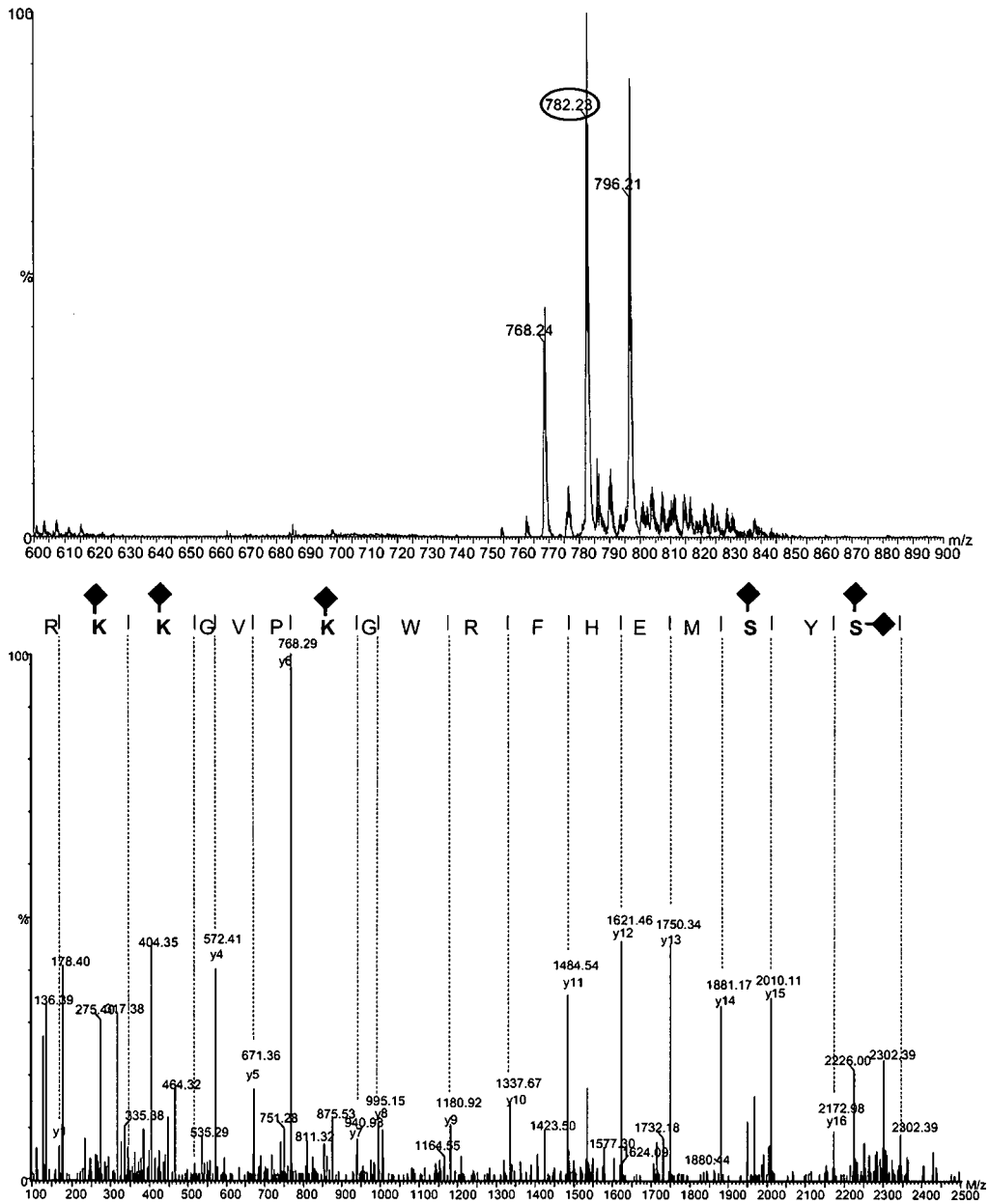


Figure 3.14. Determination of the position of acetyl groups coupled to ACTH 1-17 $[M+3H]^{3+}$ 782.26.

Sample t=1min from the acetic anhydride acetylation time course was desalted using a C18 ZipTip and diluted (1in20) using 0.1% (v/v) FA: 50% (v/v) ACN. The differentially acetylated peptide mixture was introduced into the ESI-Q-ToF by direct infusion at a rate of 0.5µl/min. From the MS spectrum (upper panel) the precursor ion at $[M+3H]^{3+}$ 782.26 was selected using the quadrupole and fragmented using a collision energy of 30%. The product ion spectra were combined and processed using MaxENT3. The resulting MS/MS spectrum was sequenced *de novo* using the Biolynx software tool (lower panel).

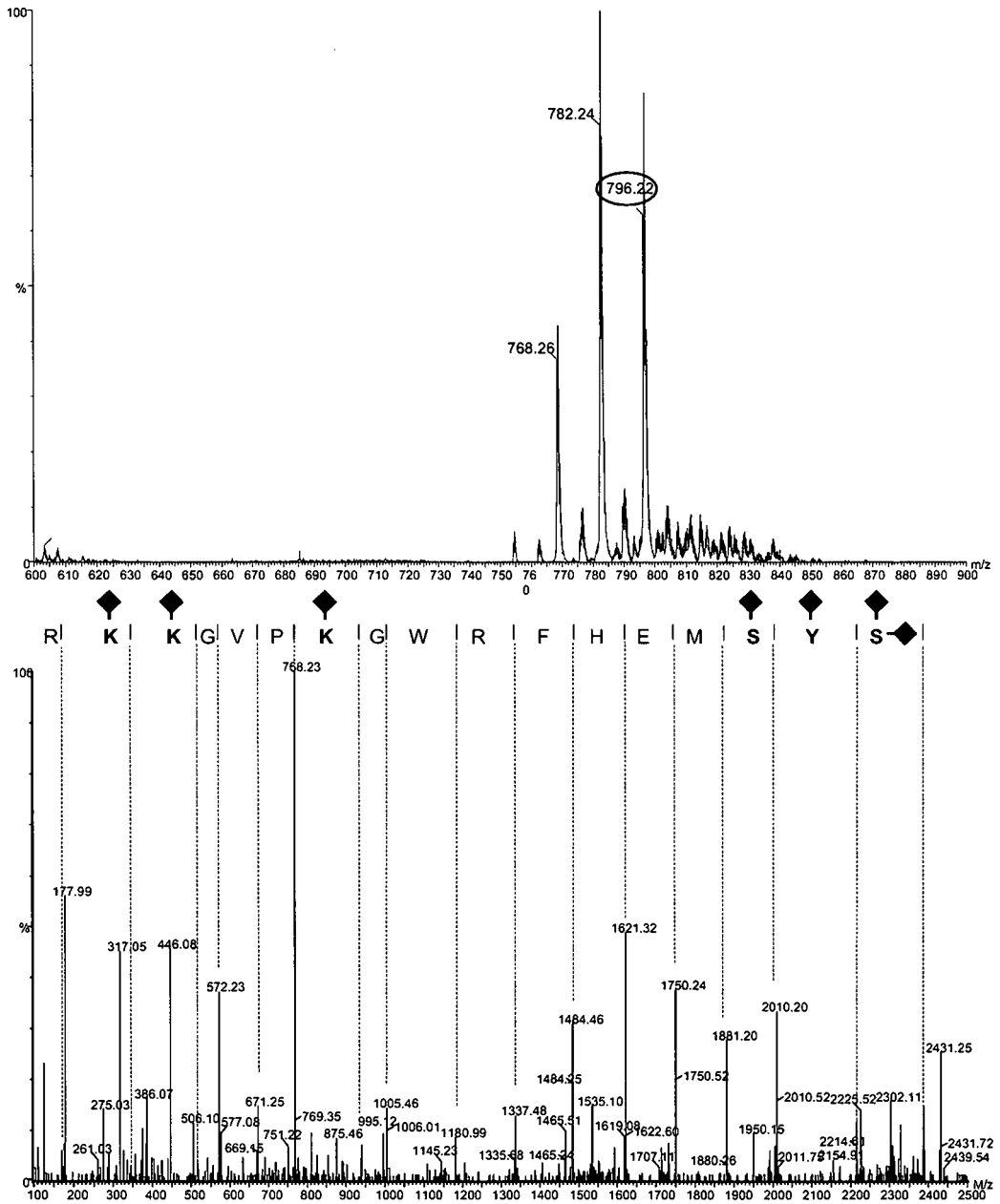


Figure 3.15. Determination of the position of acetyl groups coupled to ACTH 1-17 $[M+3H]^{3+}$ 796.26.

Sample t=120min from the acetic anhydride acetylation time course was desalted using a C18 ZipTip and diluted (1in20) in 0.1% (v/v) FA: 50% (v/v) ACN. The differentially acetylated peptide mixture was introduced into the ESI-Q-ToF by direct infusion at a rate of 0.5µl/min. From the MS spectrum (upper panel) the precursor ion at $[M+3H]^{3+}$ 796.26 was selected using the quadrupole and fragmented using a collision energy of 30%. The product ion spectra were combined and processed using MaxENT3. The resulting MS/MS spectrum was sequenced *de novo* using the Biolynx software tool (lower panel).

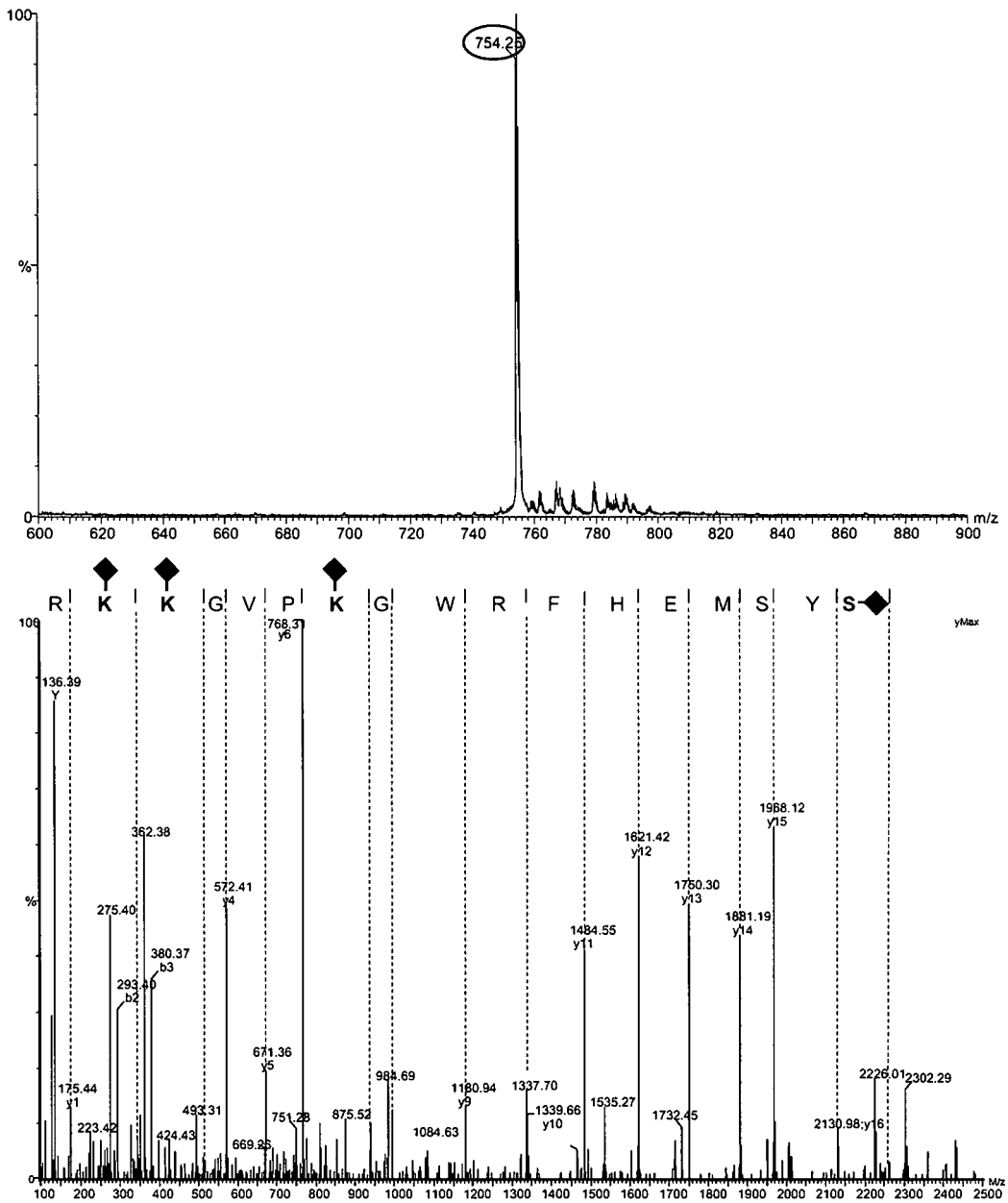


Figure 3.16. Determination of the position of acetyl groups coupled to hydroxylamine treated ACTH 1-17.

Sample t=120min (hydroxylamine treated) from the acetic anhydride acetylation time course was desalted using a C18 ZipTip and diluted (1in20) using 0.1% (v/v) FA: 50% (v/v) ACN. The differentially acetylated peptide mixture was introduced into the ESI-Q-ToF by direct infusion at a rate of 0.5µl/min. From the MS spectrum (top panel) the precursor ion at $[M+3H]^{3+}$ 754.25 was selected using the quadrupole and fragmented using a collision energy of 30%. The product ion spectra were combined and processed using MaxENT3. The resulting MS/MS spectrum was sequenced *de novo* using the Biolynx software tool (lower panel).

The behaviour observed from the two time course experiments can be explained by the strength of buffer used. Due to the fact that acetic anhydride hydrolyses rapidly in aqueous conditions to produce acetic acid, it is necessary to use a strong buffer to sustain a pH greater than pH9. The enhanced buffering capacity drives the acetylation reaction to completion faster than when a weaker buffer is used, but as a consequence leads to a greater occurrence of O-acetylation. In contrast, sulfo-NHS acetate does not hydrolyse to produce acid and can acetylate effectively in 20mM Na_2CO_3. Under these conditions it is more suitable to use sulfo-NHS acetate as the amino group acetylation agent.

3.9.3 Acetylation of a purified protein

Although it has been demonstrated that complete N-acetylation of peptides is readily achieved, the efficient acetylation of an intact protein should also be investigated. Native proteins can have complex structures that may prevent access of the reagent to specific regions of the molecule. To investigate the extent to which an intact protein can be acetylated, pyruvate kinase from rabbit muscle was derivatised using sulfo-NHS acetate.

Pyruvate kinase is a cytosolic protein and is characteristic of the soluble proteins targeted in this study. Disulfide bonds are extremely rare in cytosolic proteins, since the cytosol is generally a reducing environment (Derman and Beckwith, 1991). For this reason, it should be feasible to modify cytosolic proteins in their native state without the need for reduction and alkylation steps.

The protein was acetylated using sulfo-NHS acetate in 20mM Na_2CO_3, pH9. A comparison was made between the non-acetylated (control) tryptic digest and the acetylated protein. Samples were set up in duplicate to allow for experimental variation. Following acetylation, polymer bound Tris was added to the samples to quench residual reagent. Samples were also treated with hydroxylamine at this stage, which is required to remove O-acetylation of serine and tyrosine. Prior to digestion, all samples were subjected to a TCA precipitation and multiple ether washes to remove all traces of TCA. The protein pellets were resuspended in 50mM ammonium bicarbonate and digested overnight with trypsin. The resulting peptide mixture was analysed by MALDI-ToF MS (Figure 3.17).

In Figure 3.17a, both lysine and arginine terminated tryptic peptides have been labelled; the peptide map represents the coverage observed. The acetylated protein, when digested with trypsin, yields only arginine terminated tryptic peptides (Figure 3.17b). The

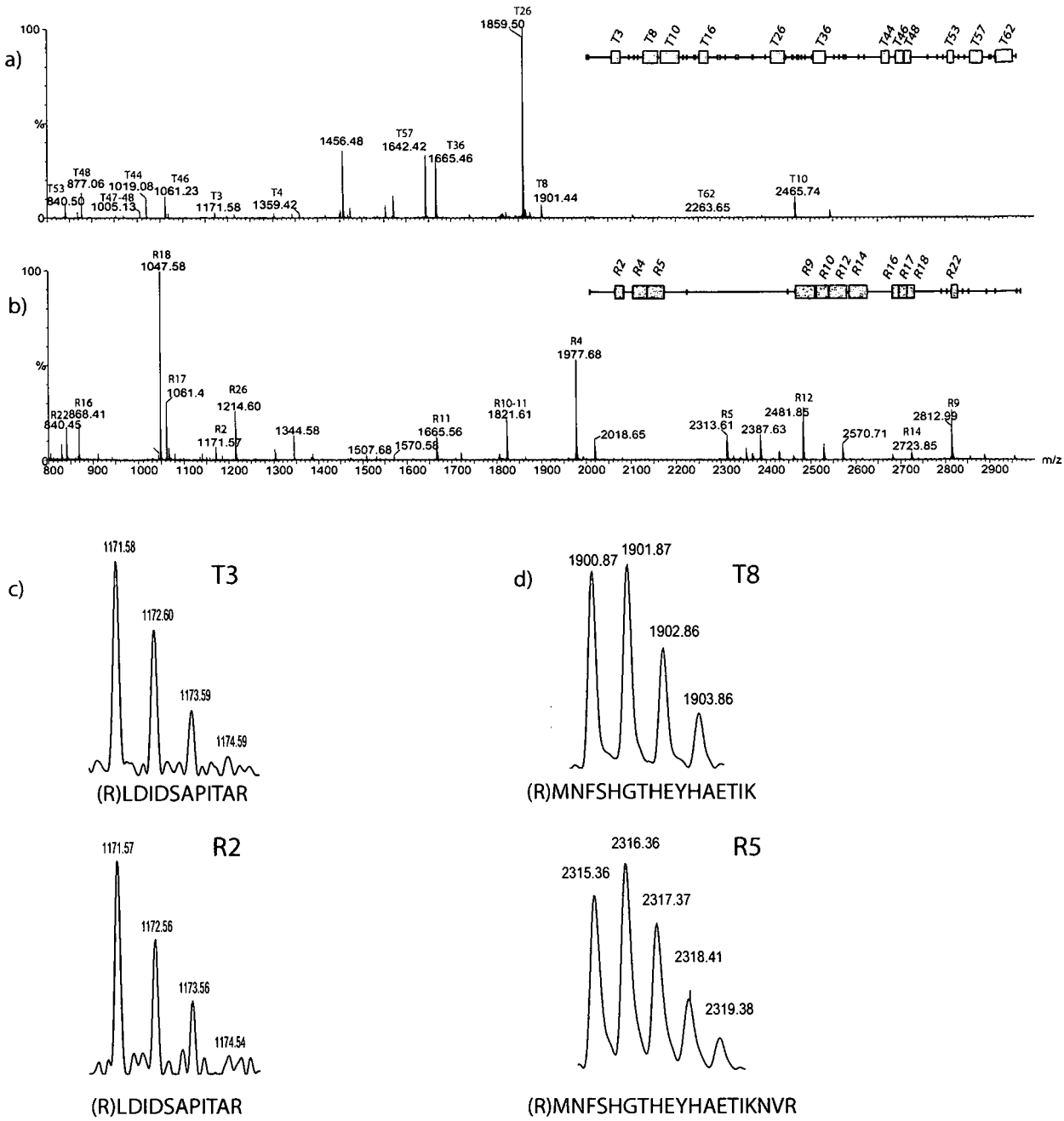


Figure 3.17. Acetylation of pyruvate kinase.

Pyruvate kinase (10µg) was acetylated by the addition of 0.1mg sulfo-NHS acetate in 20mM Na_2CO_3. The acetylated protein was then digested, alongside 10µg of unmodified protein, with trypsin (1:50 enzyme:substrate ratio) overnight. The resulting peptide mixtures, unmodified (a) and acetylated (b), were analysed by MALDI-ToF MS. The unmodified protein gave rise to both lysine and arginine terminated tryptic peptides (indicated in green of the peptide map) and the acetylated protein gave rise to arginine terminated tryptic peptides exclusively (Arg-C peptides, indicated in purple on the peptide map). Non-lysine containing tryptic peptides appeared in both digests, for example $[M+H]^+$ ion at 1171.58 (c). The absence of lysine terminated peptides from the acetylated digest, in addition to the mass shift of +42Da for lysine containing peptides (d), provides evidence that acetylation is complete.

absence of lysine cleaved tryptic peptides in Figure 3.17b infers that acetylation is complete, as proteolysis at lysine residues has been blocked by side chain modification. Peaks common to both non-acetylated and acetylated digests are the result of two successive arginine cleavages, without the presence of an internal lysine residue (Figure 3.17c). The presence of an internal lysine residue in the acetylated peptide mixture produces a mass increase of 42Da corresponding to the acetylated amino group (ε-amino group; Figure 3.17d).

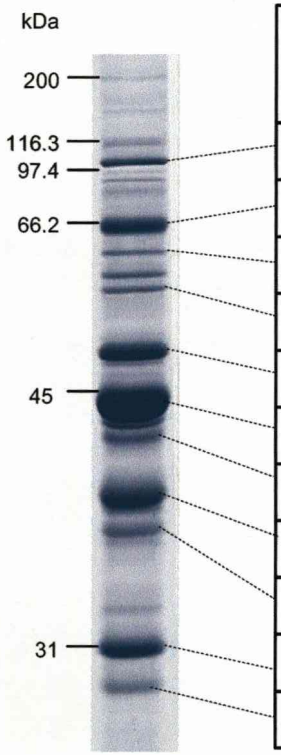
3.9.4 SDS-PAGE and PMF of mouse skeletal muscle soluble proteins

The soluble fraction of skeletal muscle is comprised mainly of glycolytic enzymes and is sufficiently enriched in a relatively small number of proteins, thus providing useful test material for the N-terminal protocol (Hayter *et al.*, 2003).

To assess the complexity and characterise the dominant protein species in mouse skeletal muscle soluble fraction, the protein mixture (15µg) was separated by 1-D SDS-PAGE and visualised using Coomassie. The resulting 1-D profile is dominated by 11 major bands, corresponding to the most abundant proteins in the sample (Table 3.3). To obtain protein identification, the major bands were excised and subjected to in-gel proteolysis using trypsin. The digested peptide mixtures were characterised by MALDI-ToF MS (Supplementary data A). The proteins were identified from their peptide mass fingerprint by manual searching against a locally implemented Mascot server against the MSDB database. Search parameters allowed a single missed tryptic cleavage, carbamidomethyl modification of cysteine (fixed) oxidation of methionine (variable), and a peptide tolerance of ±150ppm. The taxonomic space was restricted to *Mus musculus*. The proteins were all identified with high confidence and were predominantly found to be glycolytic enzymes. Figure 3.18 shows the coverage observed for each protein hit from the PMF experiment.

3.9.5 Acetylation of mouse muscle soluble fraction

The soluble fraction of mouse skeletal muscle (50µg protein) was acetylated using 1mg NHS acetate. In an attempt to decrease the number of experimental steps, the intact protein mixture was acetylated without prior reduction and alkylation. Soluble skeletal muscle proteins are mainly cytosolic proteins that have uncomplicated tertiary or quaternary structures due to lack of disulphide bridges. For this reason, it is not necessary to denature the protein by reduction in order to achieve modification. The acetylated protein mixture was treated with polymer bound Tris, which serves to remove excess acetylation reagent. The mixture was



	Protein (Accession)	Code	Mass (Da)	MOWSE score [a]	Coverage (%)	Peptides matched
1	Glycogen phosphorylase (Q9WUB3)	GP	97.7	297 (59)	57	50
2	Serum albumin (mature)(P07724)	SA	70.7	169 (59)	29	15
3	Phosphoglucomutase-1 (Q9D0F9)	PGM	61.8	178 (53)	60	35
4	Pyruvate kinase (P52480)	PK	58.4	100 (55)	26	12
5	Beta-enolase (P21550)	ENOB	47.3	122 (59)	27	10
6	Creatine kinase (P07310)	CK	43.2	256 (55)	41	19
7	Fructose-bisphosphate aldolase A (P05064)	ALDOA	39.8	141 (55)	37	12
8	Glyceraldehyde-3-phosphate dehydrogenase (P16858)	GAPDH	36.1	83 (55)	28	8
9	Phosphoglycerate mutase (O70250)	PGAM	29.9	171 (59)	42	10
10	Triose phosphate isomerase (P17751)	TPI	27.0	151 (55)	35	8
11	Adenylate kinase isoenzyme 1 (Q9R0Y5)	AK	21.5	71 (55)	36	6

Table 3.3. Identification of mouse skeletal muscle proteins using PMF.
Protein bands obtained from 1-D SDS-PAGE separations of mouse skeletal muscle soluble fraction were excised and subjected to in-gel digestion with trypsin. The resulting peptides were analysed by MALDI-ToF MS and proteins were identified using PMF searching.

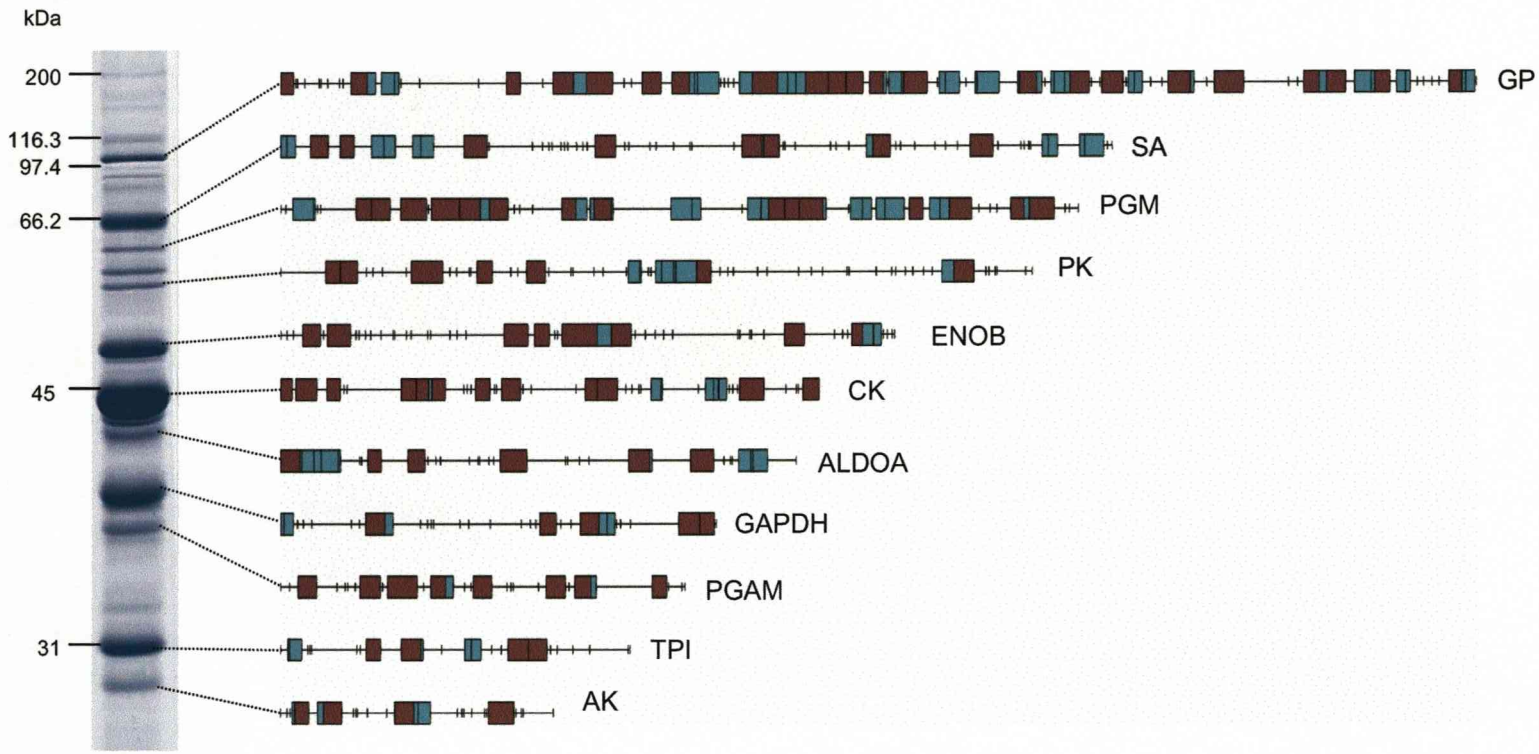


Figure 3.18. SDS-PAGE and PMF of soluble proteins from mouse skeletal muscle.
Skeletal muscle soluble fraction (20µg) was separated by 1-D SDS-PAGE and visualised using Coomassie. Spots were excised from the major bands and subjected to in-gel proteolysis using trypsin (50:1 substrate enzyme ratio). The peptides were analysed by MALDI-ToF MS and proteins identified by PMF (information on protein matches is contained in Table 3.3). The peptide coverage maps represent matched peptides (limit tryptic peptides are represented in purple, peptide matches with one missed cleavage are represented in green).

precipitated using TCA and the protein pellet washed in ether to ensure removal of residual acid, prior to overnight proteolysis. A control digest of mouse skeletal muscle soluble fraction was also prepared to compare the unmodified and acetylated digestion products. Small amounts (5µl) of the resulting peptide mixtures were desalted using a C18 ZipTip and analysed by MALDI-ToF MS. A further aliquot (1µl) of the unmodified and acetylated peptide mixture was diluted (1 in 200) and 10µl of each dilution was subjected to LC-MS/MS analysis on the LTQ ion trap instrument.

Figure 3.19 shows the peptide profiles for the unmodified (a) and acetylated (b) tryptic peptides. The MALDI-ToF spectra generated for both the unmodified and modified tryptic digests represent the most abundant peptides in the sample. Peptides generated from the acetylated protein mixture appear as arginine terminated tryptic peptides, as the lysine residues are blocked by acetylation and are resistant to proteolysis. The data generated from LC-MS/MS analysis was used to assign peptide identifications and annotate the most abundant ions in the MALDI-ToF spectra. The major ions in both spectra can be assigned to abundant proteins in skeletal muscle (seen in Figure 3.18).

MALDI-ToF analysis at this stage of the protocol serves as a 'check point' in which the effectiveness of the acetylation step can be monitored, provided a few key peptides can be identified. The precise level of acetylation is difficult to determine, however, the observation of arginine terminated, as opposed to lysine terminated, tryptic peptides provide a good indication that a sufficient degree of modification has occurred. The N-terminal isolation protocol is a "self-cleaning" method; as a result, the presence of a small amount of unmodified peptides would not have a detrimental effect on the overall procedure, as these unmodified peptides would be removed in later stages of the protocol (biotinylation of peptides/ removal of peptides with streptavidin/NHS-Sepharose). A low level of acetylation, however, will result in a reduced yield of N-terminal peptides.

3.9.6 Biotinylation of acetylated digest of mouse skeletal muscle proteins

The remaining peptide mixture produced from digestion of acetylated mouse skeletal muscle proteins (50µg), was biotinylated by the addition of 1mg NHS-biotin. The subsequent peptide mixture consisted of biotinylated internal and C-terminal peptides and non-biotinylated N-terminal peptides. Following the biotinylation reaction, the modified peptide mixture was desalted using a C18 ZipTip and subjected to MALDI-ToF analysis. Figure 3.20 shows the spectra obtained from the acetylated digest (a) and the biotinylated form of the peptide mixture

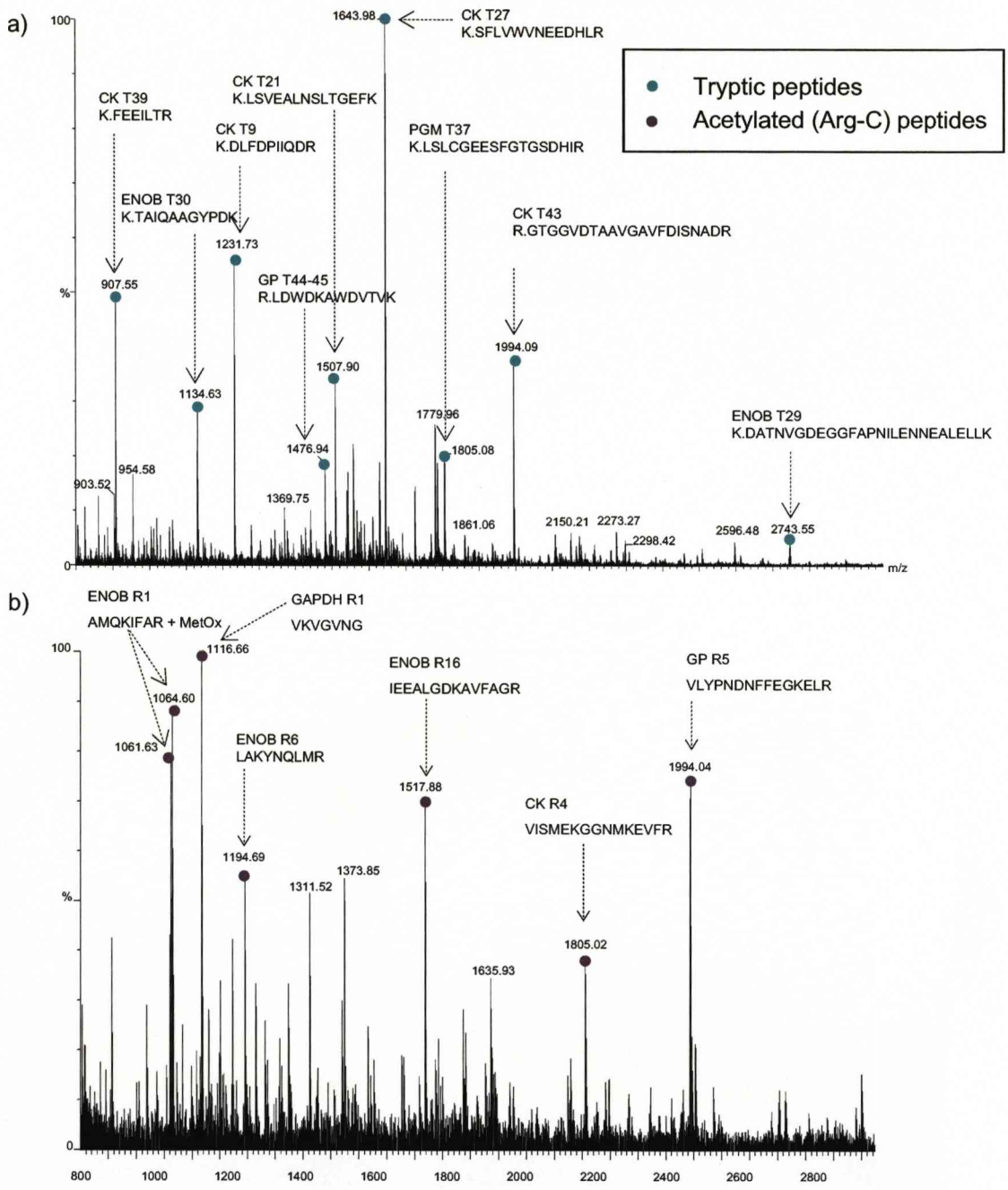


Figure 3.19. Comparison of tryptic digests from unmodified and acetylated mouse skeletal muscle soluble fraction.

Unmodified (a) and acetylated (b) mouse skeletal muscle soluble proteins were digested with trypsin (50:1 substrate enzyme ratio) at 37°C overnight. Peptides were analysed by MALDI-ToF MS using a laser energy of 30%. Peptide ions in (a) were representative of the most abundant proteins in mouse skeletal muscle. Tryptic peptides identified by LC-MS/MS on the LTQ ion trap instrument were matched to the most intense ions in the spectrum. The acetylated protein mixture (b), when digested with trypsin, yielded arginine terminated peptides only, containing a mass shift of +42Da for each internal lysine. Peptides from the major proteins in the sample were matched to the most intense ions. The absence of ions corresponding to lysine terminated tryptic peptides coupled with the observation of a +42Da increase for each internal lysine in (b) is sufficient to confirm peptide acetylation.

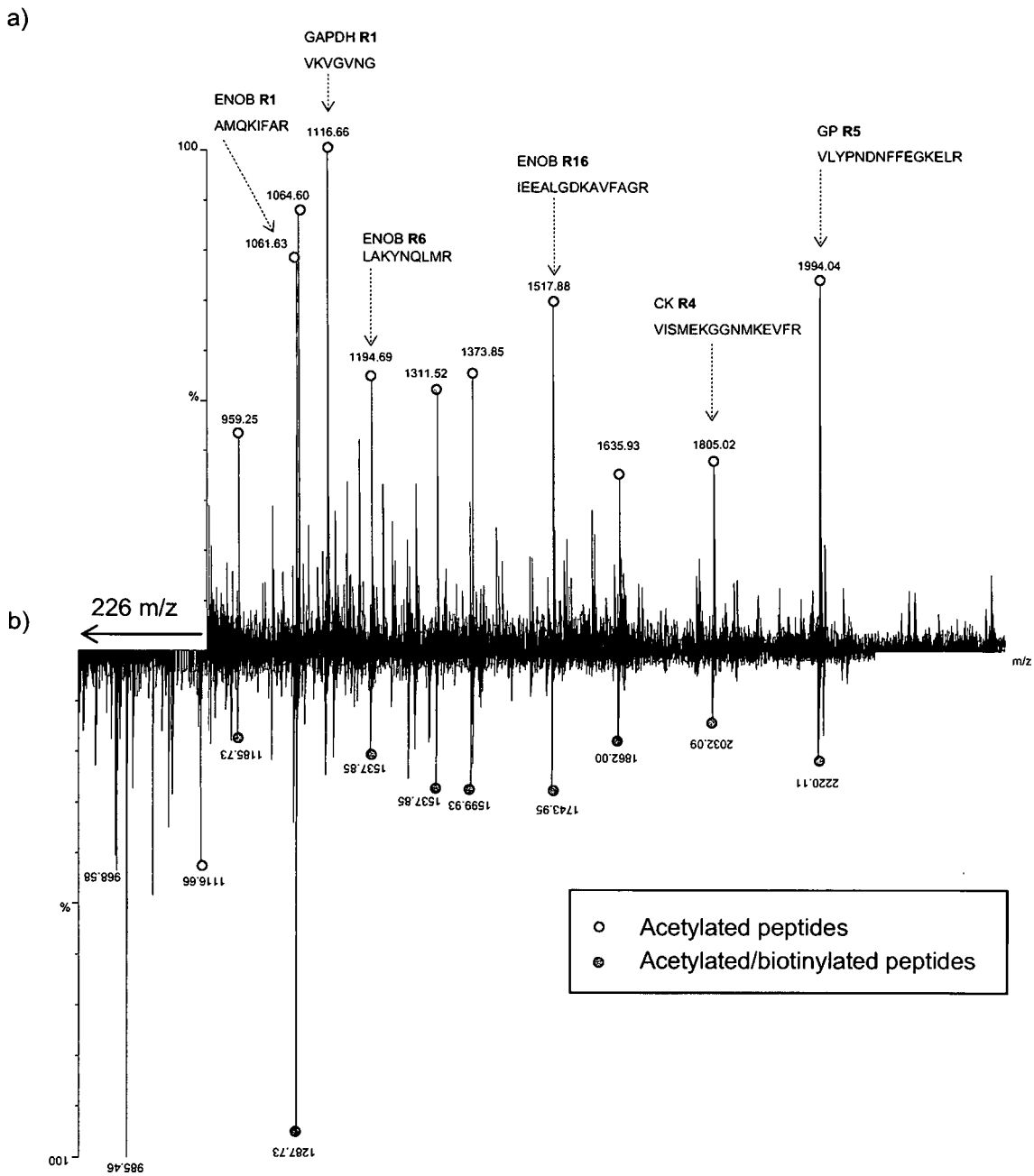


Figure 3.20. Comparison of acetylated and acetylated/biotinylated peptides from mouse skeletal muscle soluble fraction.
Acetylated (a) and acetylated/biotinylated (b) peptides were analysed by MALDI-ToF MS using a laser energy of 30%. The majority of the ions exhibited a mass shift of +226Da indicating the addition of a biotin group.

(b). All internal peptides contain a free α-amino group and react with the NHS-biotin undergoing a mass shift of +226Da. In contrast to the acetylation step, it is important that biotinylation of peptides is complete. If any internal peptides remain unmodified they will not be retained by the streptavidin Sepharose and will pass through along with the N-terminal peptides. Tryptic peptides are, in general, easier to modify than intact proteins due to their size and accessibility. A 100-fold excess of NHS-biotin in a controlled reaction environment (pH7.5) is sufficient to ensure complete biotinylation.

3.9.7 Streptavidin purification of internal peptides

To separate internal from N-terminal peptides, the biotinylated peptide mixture was subjected to affinity purification by streptavidin Sepharose. The Sepharose (20µl) was washed three times with binding buffer (20mM Na_2HPO_4, 0.15M NaCl, pH7.5) before adding the starting material equivalent of 1µg desalted peptides (made up to 20µl in binding buffer). The flow through was retained and the column was washed with a further 10µl of binding buffer. The unbound material containing the purified N-terminal peptides was pooled. A small amount of flow through was desalted using a C18 ZipTip and analysed by MALDI-ToF MS and the remaining material was subjected to LC-M/SMS analysis on the LTQ ion trap instrument. Figure 3.21 represents the MALDI-ToF spectrum of the purified N-terminal peptides. The most notable observation is the high level of simplification gained by the removal of internal peptides from the complex mixture. The N-terminal MALDI-ToF spectrum contains around 10 major peaks, in contrast with the spectrum from the in-solution digests of mouse muscle, which contain in excess of 50 intense peaks (Figure 3.19). Table 3.4 lists the most abundant proteins in mouse skeletal muscle soluble fraction (as seen on gel in Figure 3.19), along with the predicted N-terminal "Arg-C" sequences and $[M+H]^+$ values. Three major ions present in the MALDI-ToF spectrum (Figure 3.21), correspond to N-terminal peptides from beta enolase (m/z 1048.59) and GAPDH (m/z 1116.66). The ion at m/z 1060.59 corresponds to the methionine oxidised form of the beta enolase peptide at m/z 1048.59. Lower intensity signals are also present for the N-termini of annexin (m/z 1191.69), serum albumin (m/z 1261.68) and glycogen phosphorylase (m/z 1285.70). The N-terminal peptides from the two proteins fructose-bisphosphate aldolase and phosphoglucomutase, both have similar m/z values (2553.22 and 2553.38 respectively). The resolution of the MALDI-ToF instrument is not capable of distinguishing between these two m/z values. The remaining low abundant ions in the spectrum remain unidentified.

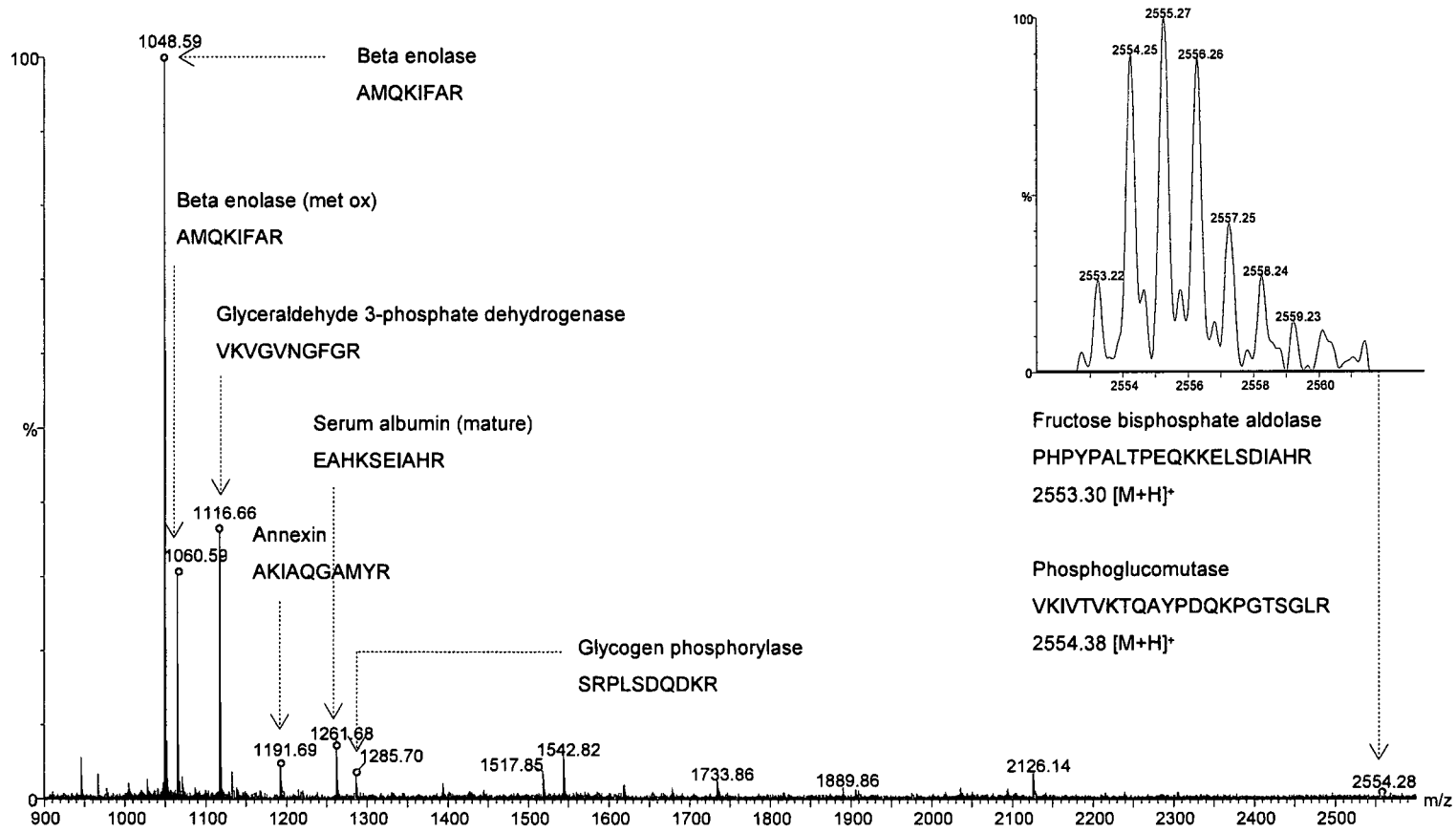


Figure 3.21. N-terminal purification of mouse skeletal muscle soluble proteins.
The purified N-terminal peptides were desalted using a C18 ZipTip and analysed by MALDI-ToF MS using a laser energy of 30%. The major ions can all be assigned to predicted N-terminal peptides from Table 3.4. In addition to the predicted N-terminal peptides, LC-MS/MS resulted in the identification of more proteins including annexin, which can also be seen in the spectrum. The N-terminal peptides of fructose bisphosphate aldolase and phosphoglucomutase are present but at much lower intensities than the other peptide ions. The unlabelled peaks remain unidentified.

Protein	N-terminal sequence	[M+H]+	Acetylation sites	Acetylated [M+H]+
Gycogen phosphorylase	SRPLSDQDKR	1200.62	2	1284.63
Serum albumin	EAHKSEIAHR	1176.61	2	1260.62
Phosphglucomutase	VKIVTVKTQAYPDQKPGTSGLR	2385.34	4	2553.38
Pyruvate kinase	PKPHSEAGTAFIQTQQLHAAMA DTFLEHMCR	3465.65	2	3550.66
β-enolase	AMQKIFAR	964.54	2	1048.55
Creatine kinase	PFGNTHNKFKLNYKPQEEYPDLSKHNNHMAKVLTPDLYNKLR	5038.57	7	5332.64
Fructose bisphosphate aldolase	PHPYPALTPEQKKELSDIAHR	2426.27	3	2552.30
Glyceraldehyde 3 -phosphate dehydrogenase	VKVGVNGFGR	1032.59	2	1116.61
Lactate dehydrogenase	ATLKDHLIHNVHKEEHAHAHNKISVVGVGAVGMACAISILM KD LADELTLVDVVEDKLKGEMLDLQHGSLFLKTPKIISGKD YSVT AHSKLVIVTAGAR	10636.70	11	11098.81
Phosphoglycerate mutase	ATHR	483.25	1	515.27
Triose phosphate isomerase	APTR	443.24	1	485.25
Adenylate kinase	EEKLKKAKIIFVVGGPGSGKGTQCEKI VQKYGYTHLSTGDLLR	4675.55	8	5011.63

Table 3.4. Predicted N-terminal peptides from the major proteins in mouse skeletal muscle soluble fraction.
Sequences were retrieved from the SwissProt database for the most abundant proteins in mouse muscle soluble fraction. $[M+H]^+$ values of the unmodified and acetylated forms of the N-terminal Arg-C peptides were determined using the protein/peptide editor tool in MassLynx.

LC-MS/MS analysis of the mouse muscle N-terminal sample identified a total of five N-terminal "Arg-C" peptides from mouse skeletal muscle (Table 3.5). The relatively low number of hits can be attributed partly to the nature of skeletal muscle soluble proteins, which is dominated by approximately 10-20 major proteins (predominantly glycolytic enzymes; Hayter *et al.*, 2003; Doherty *et al.*, 2004). This preparation has the advantage of generating relatively simple spectra and provides a valuable test system with which to establish the method. The low number of hits is also due to the reduced yield of N-terminal peptides generated by the biotin/streptavidin method.

3.9.8 Removal of internal peptides by NHS-activated Sepharose

The acetylated peptide mixture was diluted 1:1 in binding buffer and incubated with NHS-Sepharose for 4h at room temperature (with turning), then again in a fresh aliquot of Sepharose overnight at 4°C. The supernatant containing the N-terminal peptides was removed from the Sepharose. A small amount (5µl) of N-terminal peptides was desalted using a C18 ZipTip. The desalted peptide preparation was analysed by MALDI-ToF MS. The remaining material (100µl) was split into 10µl aliquots, which were either subjected to LC-MS/SMS analysis on the LTQ ion trap instrument, or stored at -20°C for later use. Figure 3.22 shows a comparison of the MALDI-ToF spectrum observed for the N-terminal peptides generated by both methods. There appears to be no difference in the major peaks seen in both spectra, which indicates that both methods are equally effective for separating non-acetylated internal from acetylated N-terminal peptides. The LC-MS/MS analysis from the N-terminal peptides generated in the NHS-Sepharose method resulted in the identification of 50 N-terminal peptides (Table 3.6) representing a 10-fold increase in identifications compared with the biotin method. Out of these 50 peptides, 41 proteins had undergone NME and four SP removal. The remaining five peptides were unmodified.

The gain in peptide identification can be attributed to the increased yield of product obtained in the NHS-Sepharose method, which is a direct result of fewer analytical steps. Due to the increased yield the NHS-Sepharose method for N-terminal purification will be the method adopted for use on more complex samples.

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score
1	Beta-enolase (P21550)	1047.55	AMQKIFAR	M	102
2	Glyceraldehyde-3-phosphate dehydrogenase (P16858)	1115.61	VKVGVNGFGR	M	150
3	Bisphosphoglycerate mutase (P15327)	1232.76	SKHKLIILR	M	54
4	Serum albumin (mature)(P07724)	1260.62	EAHKSEIAHR	SP	475
5	Glycogen phosphorylase, muscle form (Q9WUB3)	1326.65	SRPLSDQDKR	M	32
6	Fructose-1,6-bisphosphatase isozyme 2 (P70695)	1823.87	TDRSPFETDMLTLTR	M	47
7	Fructose-bisphosphate aldolase A (P05064)	2552.30	PHPYPALTPEQKKELSDIAHR	M	29

Table 3.5. Identification of mouse skeletal muscle proteins by LC-MS/MS analysis of N-terminal peptides generated using the NHS-Biotin method.
The N-terminal peptide preparation of mouse skeletal muscle soluble fraction was analysed by LC-MS/MS on the LTQ ion trap instrument using a three hour RP gradient. MS/MS data was used to search the mouse N-terminal database using the MASCOT search engine. The taxonomy was restricted to *Mus musculus;* fixed modifications: N-terminal acetylation and lysine acetylation; variable modification: oxidation of methionine; protease: Arg-C; missed cleavages: 1; peptide tolerance: 1.5Da; MS/MS tolerance: 0.6Da; instrument: ESI-TRAP; peptide charge: 1+, 2+ and 3+. Protein identifications with a Mowse score greater than 20 were accepted as confident identifications.

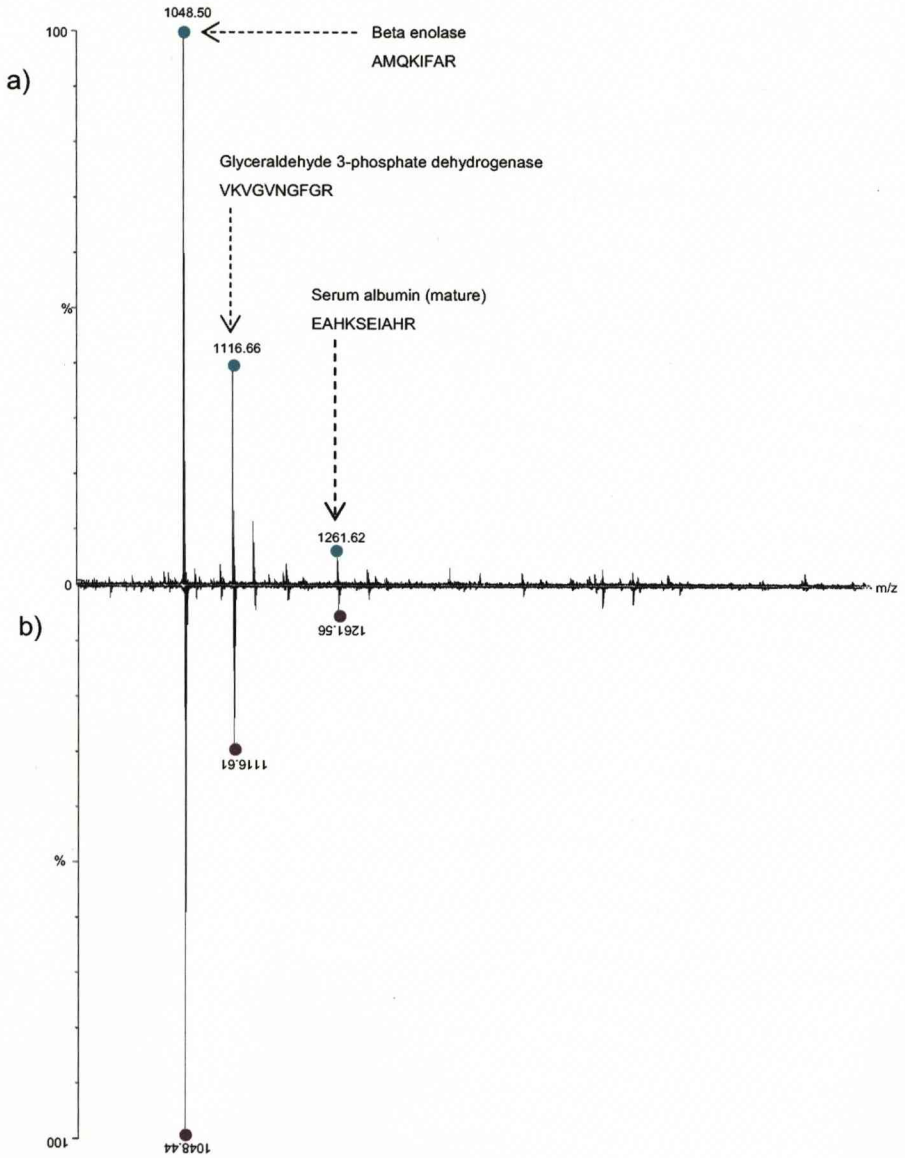


Figure 3.22. Comparison of NHS-biotin and NHS-Sepharose methods for N-terminal purification.
MALDI-ToF spectra from the N-terminal peptide preparations generated using the NHS-biotin (a) and the NHS-Sepharose methodologies (b), are displayed for comparative purposes.

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score
1	40S ribosomal protein S12 (P63323)	701.32	AEEGIAR	M	34
2	Nuclear localisation protein 4 (P60670)	842.49	AESIIIR	M	21
3	Ferritin heavy chain (P09518)	987.50	TTASPSQVR	M	29
4	Ribonuclease P protein subunit p30 (O88796)	1032.52	AAFADLDLR	M	42
5	Beta-enolase (P21550)	1047.55	AMQKIFAR	M	210
6	Voltage-dependent calcium channel (O70578)	1059.60	SQTKTAKVR	M	20
7	Homeobox protein Hox-C13 (P50207)	1078.61	TTSLLLHPR	M	25
8	Bisphosphate 3'-nucleotidase 1 (Q9ŒS1)	1084.53	ASSHTVLMR	M	56
9	S-formylglutathione hydrolase (Q9R0P3)	1099.60	ALKQISSNR	M	41
10	Glyceraldehyde-3-phosphate dehydrogenase (P16858)	1115.61	VKVGVNGFGR	M	190
11	Dimethylarginine dimethylaminohydrolase (Q9CWS0)	1152.57	AGLGHPSAFGR	M	41
12	14-3-3 protein eta (P68510)	1155.60	GDREQLLQR	M	87
13	Annexin A6 (P14842)	1191.61	AKIAQGAMYR	M	98
14	Mitochondrial 28S ribosomal protein S21 (P58059)	1208.70	AKHLKFIAR	M	50
15	Bisphosphoglycerate mutase (P15327)	1232.76	SKHKLIILR	M	89
16	Serum albumin (mature)(P07724)	1260.62	EAHKSEIAHR	M	150
17	40S ribosomal protein S15 (P62843)	1282.69	AEVEQKKKR	M	54
18	Glutathione S-transferase Mu 1 (P10649)	1289.66	PMILGYWNVR	M	32
19	Glutathione S-transferase Mu 2 (P15626)	1292.62	PMTLGYWDIR	M	62
20	Glycogen phosphorylase (Q9WŒ3)	1326.65	SRPLSDQDKR	M	78
21	Macrophage migration inhibitory factor (P34884)	1328.69	PMFIVNTNVPR	M	23
22	Peptidyl-prolyl cis-trans isomerase (P26883)	1355.67	GVQVETISPGDGR	M	47
23	Glutathione S-transferase P1 (P19157)	1392.74	PPYTIVYFPVR	M	50
24	14-3-3 protein gamma (P61982)	1424.77	DREQLVQKAR	M	71
25	Fatty acid-binding protein (Q05816)	1427.74	ASLKDLEGKWR	M	29
26	Tubulin polymerisation-promoting protein (Q9CRB6)	1507.70	AASTDIAGLEESFR	M	20
27	Calseqestrin-1 (mature) (O09165)	1667.70	EDGLDFPEYDGVDR	SP	40
28	Acyl-CoA-binding protein (P31786)	1782.33	SQAEFDKAAEEVKR	M	48
29	aspartate-beta-hydroxylase (CAM26685)	1800.85	AEDKEAKHGGHKNGR	M	75
30	Fructose-1,6-bisphosphatase isoŒme 2 (P70695)	1823.87	TDRSPFETDMLTLTR	M	89
31	Myosin-1 (Q5SX40)	184.850	SSDAEMAVFGEAAPYLR	M	56
32	Cytosolic non-specific dipeptidase (Q9D1A2)	1979.96	SALKAVFQYIDENQDR	M	30
33	Peptidyl-prolyl cis-trans isomerase A (P17742)	2046.98	VNPTVFFDITADDEPLGR	M	51
34	Phosphoglycerate kinase 1 (P09411)	2166.19	SLSNKLTLDKLDVKGKR	M	96
35	Peroxiredoxin-6 (O08709)	2183.09	PGGLLLGDEAPNFEANTTIGR	M	75
36	Cofilin-1 (P18760)	2217.15	ASGVAVSDGVIKVFNDMKVR	M	102
37	14-3-3 protein eta/delta (P63101)	2268.14	MDKNELVQKAKLAEQAER		136
38	Myosin light chain 3 (P05978)	2303.11	SFSADQIADFKEAFLLFDR	M	23
39	Cofilin-2 (P45591)	2346.19	ASGVTVNDEVIKVFNDMKVR	M	48
40	14-3-3 protein epsilon (P62259)	2363.11	MDDREDLVYQAKLAEQAER		69
41	Calreticulin (mature) (P14211)	2369.10	DPAIYFKEQFLDGDAWTNR	SP	48
42	Histidine triad nucleotide-binding protein 1 (P70349)	2523.33	ADEIAKAQVAQPGGDTIFGKIIR	M	21
43	Fructose-bisphosphate aldolase A (P05064)	2552.30	PHPYPALTPEQKKELSDIAHR	M	90
44	Phosphoglucomutase-1 (Q9D0F9)	2553.38	VKIVTVKTQAYPDQKPGTSGLR	M	87
45	Tropomyosin-1 alpha chain (P58771)	2770.41	MDAIKKKMQMLKLDKENALDR		69
46	Adenylate kinase isoenŒme 1 (Q9R0Y5)	2823.51	GKKLSAIMEKGELVPLDTVLDMLR	M	49
47	Actin, cytoplasmic 2 (P63260)	2861.28	EEEIAALVIDNGSGMCKAGFAGDDAPR		68
48	Hemoglobin subunit alpha (P01942)	3380.66	VLSGEDKSNIKAAWGKIGGHGAEYGAEALER		47
49	Alpha-1-antitrypsin 1-2 (mature) (P22599)	3748.73	EDVQETDTSQKDQSPASHEIATNLGDFAISLYR	SP	40
50	Alpha-1-antitrypsin 1-4 (mature) (Q00897)	3385.55	EDVQETDTSQKDQSPASHEIATNLGDFALR	SP	59

Table 3.6. Identification of mouse skeletal muscle proteins by LC-MS/MS analysis of N-terminal peptides generated using the NHS-Sepharose method.
The N-terminal peptide preparation of mouse skeletal muscle soluble fraction was analysed by LC-MS/MS using a three hour RP gradient. MS/MS data was used to search the mouse N-terminal database using the MASCOT search engine. The taxonomy was restricted to *Mus musculus;* fixed modifications: N-terminal acetylation and lysine acetylation; variable modification: oxidation of methionine; protease: Arg-C; missed cleavages: 1; peptide tolerance: 1.5Da, MS/MS tolerance: 0.6Da, instrument: ESI-TRAP, peptide charge: 1+, 2+ and 3+. Protein identifications with a Mowse score greater than 20 were accepted as confident identifications.

3.9.9 Atypical isotope distribution of the GAPDH N-terminal peptide

When analysing the MALDI-ToF spectrum of the mouse skeletal muscle N-terminal peptide mix, the peak representing the GAPDH N-terminal peptide exhibited an abnormal isotope distribution (Figure 3.23a). The same isotope distribution is also observed for the unmodified form of the tryptic peptide obtained from an in-gel digest of GAPDH the mouse skeletal muscle sample (Figure 3.23b). In addition, analysis of soluble proteins from skeletal muscle over a range of species reveals the same isotope distribution for the protein (data not shown).

The mass isotopmer envelope is consistent with the analyte being a mixture of two peptides: one with a monoisotopic mass of 1031.6Da and one with the monoisotopic mass of 1032.6Da. The higher m/z value could have arisen from contamination; however, since the peptide contains an asparagine residue, the most probable explanation for the 1Da mass increase is deamidation of asparagine to aspartic acid. A combination of the acid and amide form of the peptide would account for the atypical isotope pattern observed.

To validate this hypothesis, the in-gel digest of mouse muscle GAPDH (taken from the SDS-PAGE analysis in Figure 3.18) was subjected to methyl esterification. This reaction results in the addition of a methyl group (+14Da) to each acidic residue in the peptide. In the case of the asparagine containing peptide this should result in one methyl esterification to the carboxyl group on the C-terminal of the peptide. In the case of the aspartic acid peptide this should result in two methyl esterifications, one on the carboxyl group on the C-terminal of the peptide and an additional methyl esterification on the carboxyl group of the acidic residue.

In the MALDI-ToF spectrum obtained after methyl esterification the ions in the 1032-1036 m/z region disappeared, and two new ions appeared, the first representing the peptide with the addition of one methyl group (m/z 1032.58 + 14.03 = 1046.61) and the second representing the peptide with the addition of two methyl groups (m/z 1033.58 + 28.06 = 1061.64; Figure 3.23c). This experiment confirms that the atypical isotope pattern observed for the N-terminal peptide of GAPDH is indeed the result of deamidation. What is not clear from this, however, is whether this deamidation event occurs *in vivo* or is an artifact of sample preparation and processing. Further analysis (monitored proteolysis and absolute quantification of proteolytic products) showed that this modification occurs post-proteolysis (*in vitro*), due to relieved constraint of the peptide backbone trajectory, which permits the deamidation reaction to take place (Rivers *et al.*, 2008).

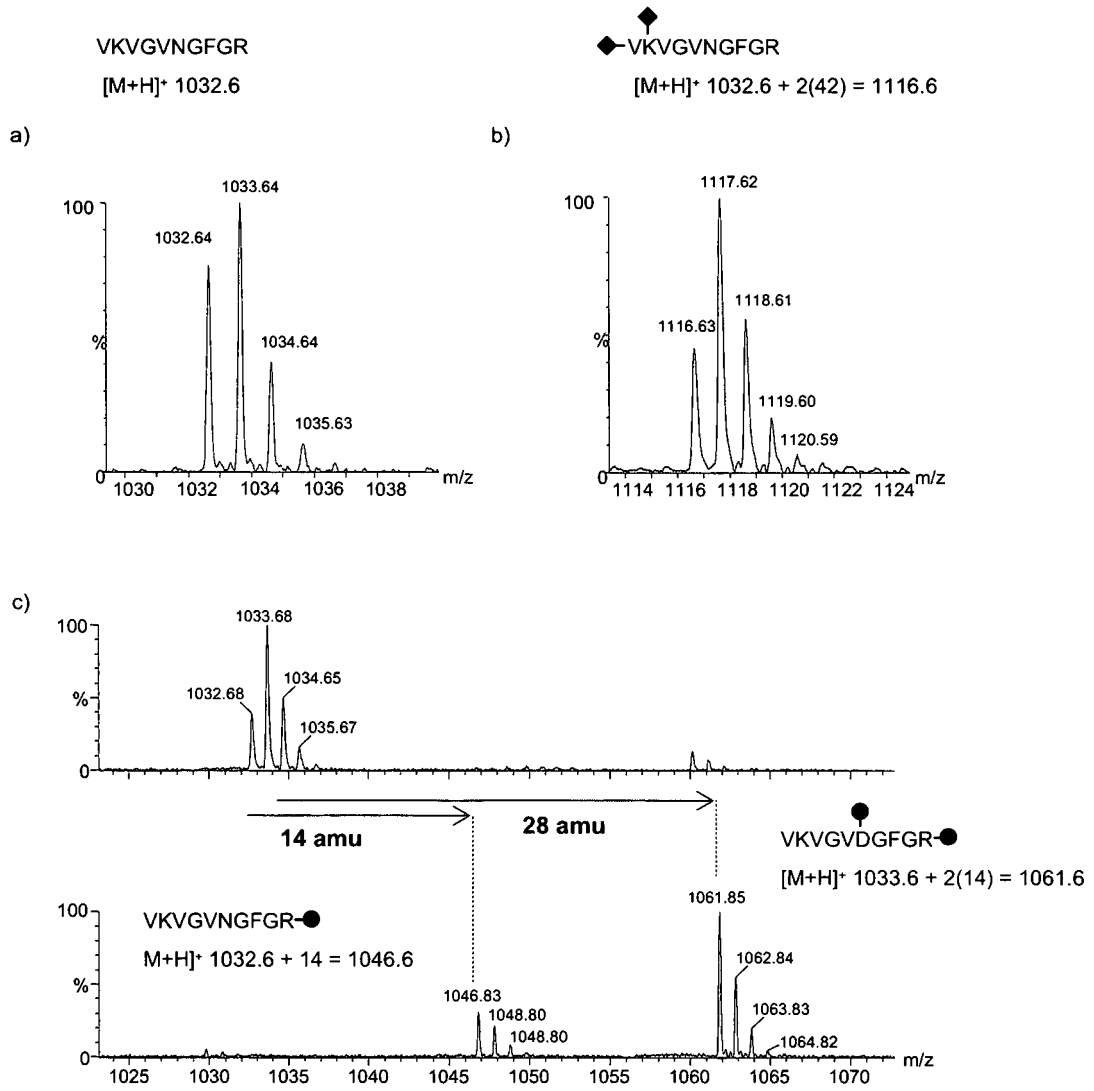


Figure 3.23. Atypical isotope distribution of the GAPDH N-terminal.
Both the unmodified (a) and N-acetylated (b) forms of the GAPDH tryptic peptide exhibit an atypical isotope distribution when analysed by MALDI-ToF MS. When reacted with acetyl chloride and methanol (esterification), the two overlapping forms of the peptide are resolved into two distinct reaction products (c). The asparagine containing peptide undergoes the addition of one methyl group (+14Da) and the aspartic acid containing peptide undergoes the addition of two methyl (+28Da) groups.

3.9.10 Global analysis of complex proteomes using positional proteomics

The NHS-Sepharose method for N-terminal enrichment was applied to three complex biological samples: mouse liver, *S. cerevisiae* and *E. coli* cell lysate. In all cases, only the soluble protein fraction was analysed. To allow for experimental variation, three separate N-terminal preparations were made and each preparation was subjected to LC-MS/MS analysis on the LTQ ion trap instrument using the extended gradient. Three separate LC-MS/MS experiments were performed for each N-terminal preparation. The experimental strategy is summarised in Figure 3.24.

Each of the acetylated digests were initially analysed by MALDI-ToF MS to provide a comparison between the total acetylated digest and the N-terminal preparation (Figures 3.25-3.27a). Due to the complexity of the peptide mixtures generated, it is difficult to assign peptides to individual peaks in the spectra. Following N-terminal enrichment by NHS-Sepharose, the mixtures generated spectra in which all major ions can be assigned to N-terminal peptides (Figures 3.25-3.27b; MALDI-ToF peak assignments are listed in Table 3.7, 3.8 and 3.9).

Although simplified, the N-terminal mixtures generated are too complex to be analysed without separation. Each sample was subjected to three identical LC-MS/MS runs on the LTQ ion trap instrument, using the extended, three hour, RP gradient (a total of nine LC-MS/MS experiments for each proteome). The MS/MS data was processed in Sequest to generate an MGF file which was used for database searching (SwissProt and specialised N-terminal databases) through the Mascot search engine. Parameters included: a single missed tryptic cleavage; acetylation of lysine and the N-terminal amino group (fixed) oxidation of methionine and acetylation of serine (variable); peptide tolerance: 1.5Da; MS/MS tolerance: 0.6Da; instrument: ESI-TRAP; peptide charge: 1+, 2+ and 3+ and the taxonomic space was restricted to the species analysed. MS/MS matches with ion scores over the reported threshold (indicating extensive homology) were taken as significant identifications.

Additionally, MS/MS matches reported with ion scores lower than the given threshold were manually inspected, and provided a strong y or b ion series could be identified, these matches were included as significant N-terminal identification. All identifications obtained by LC-MS/MS are listed in Supplementary data B. A total of 220 were identified for mouse liver, 186 for *S. cerevisiae* and 337 for *E. coli*.

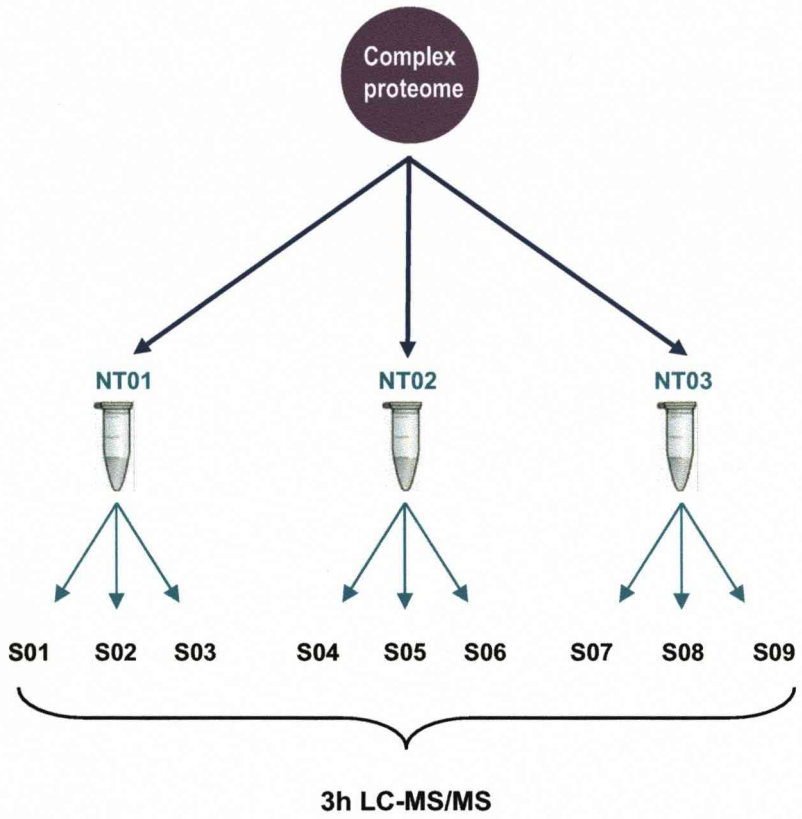


Figure 3.24. Scheme for the preparation and analysis of N-terminal peptides from three complex biological samples.
To allow for experimental variation, N-terminal preparations from each complex proteome (mouse liver, *S. cerevisiae* and *E. coli*) were generated in triplicate. Each preparation was analysed a total of three times on the LTQ ion trap instrument, using the extended three hour RP gradient (nine LC-MS/MS runs in total).

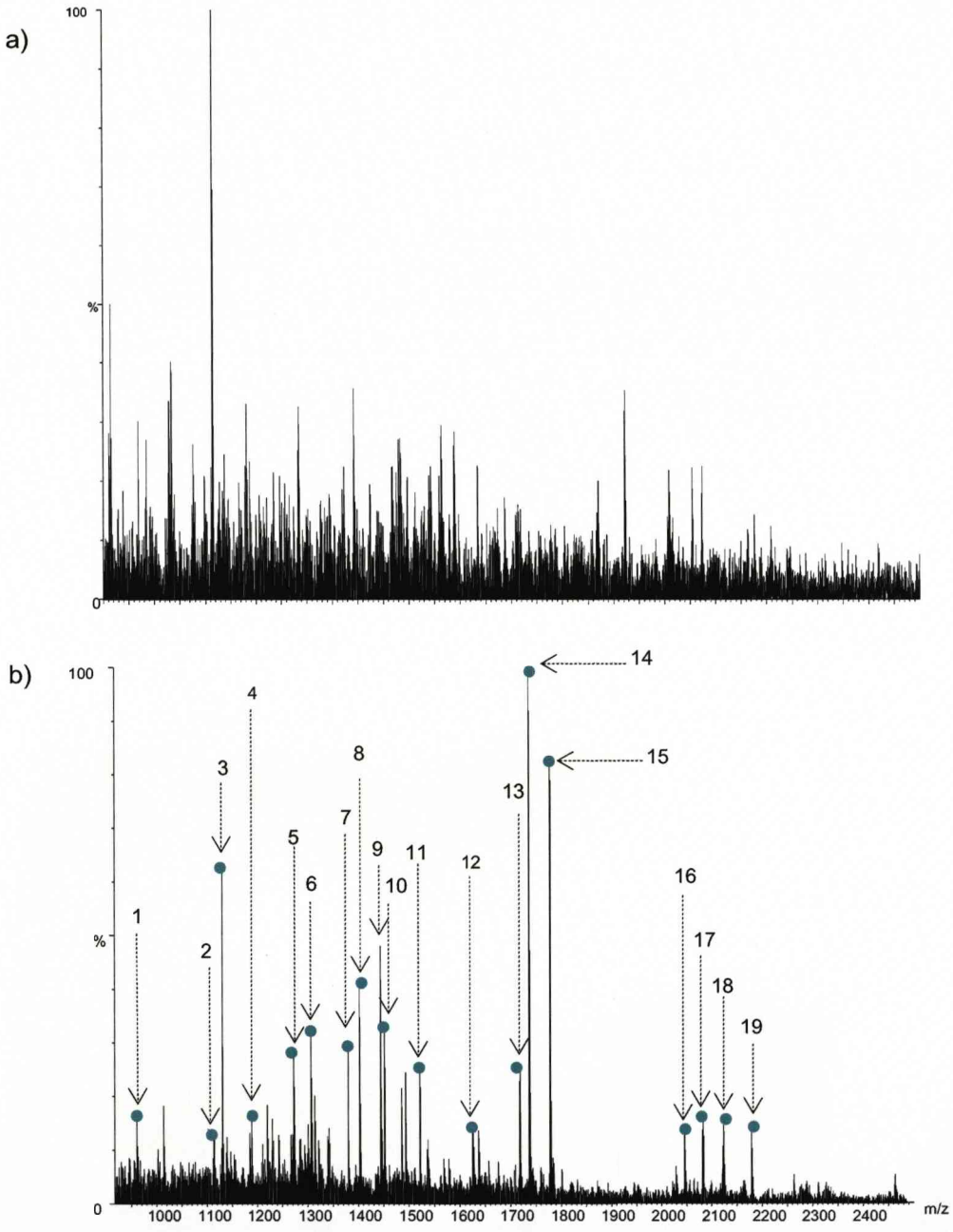


Figure 3.25. Isolated N-terminal peptides from mouse liver soluble fraction.
The acetylated digest (a) and the N-terminal preparation (b) of mouse liver proteins were analysed by MALDI-ToF MS using a laser energy of 30%. All of the major ions in the N-terminal spectra can be assigned to N-terminal peptides from mouse liver (Table 3.7 represents the labelled ions).

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score
1	60S ribosomal protein L39 (P62892)	945.47	SSHKTFR	M	58
2	Esterase D (Q9R0P3)	1099.6	ALKQISSNR	M	81
3	Glyceraldehyde-3-phosphate dehydrogenase (P16858)	1115.61	VKVGVNGFGR	M	118
4	Estradiol 17 beta-dehydrogenase 5 (P70694)	1175.56	MDSKQQTVR		125
5	Serum albumin (mature) (P07724)	1260.62	EAHKSEIAHR	SP	129
6	Glutathione S-transferase Mu 2 (P15626)	1292.62	PMTLGYWDIR	M	102
7	Sterol-4-alpha-carboxylate 3-dehydrogenase (Q9R1J0)	1370.64	MEQAVHGESKR		20
8	Glutathione S-transferase P 1 (P19157)	1392.74	PPYTIVYFPVR	M	356
9	Fatty acid-binding protein (Q05816)	1427.33	ASLKDLEGKWR	M	29
10	Glutathione S-transferase Yc (P30115)	1442.73	AGKPVLHYFDGR	M	161
11	D-dopachrome tautomerase (O35215)	1513.78	PFVELETNLPASR	M	349
12	Liver carboxylesterase 31 (mature) (Q63880)	1619.85	PKVTQPEVDTPLGR	SP	100
13	Liver fructose 1, 6 Bisphoshate (Q9QXD6)	1713.83	ANHAPFETDISTLTR	M	304
14	Acyl-CoA-binding protein (P31786)	1732.83	SQAEFDKAAEEKVR	M	391
15	Betaine–homocysteine S-methyltransferase (O35490)	1775.03	APVAGKKAKKGILER	M	678
16	Peptidyl-prolyl cis-trans isomerase A (P17742)	2046.99	VNPTVFFDITADDEPLGR	M	243
17	Sorbitol dehydrogenase (Q64442)	2084.1	AAPAKGENLSLVVHGPGDIR	M	85
18	Phosphoglycerate kinase 1 (P09411)	2124.18	SLSNKLTLDKLDVKGKR	M	98
19	Peroxiredoxin 6 (O08709)	2183.06	PGGLLLGDEAPNFEANTTIGR	M	247

Table 3.7. N-terminal peptides from soluble mouse liver proteins, observed by MALDI-ToF MS.

The N-terminal peptide preparation of mouse skeletal muscle soluble fraction was analysed by LC-MS/MS using a three hour RP gradient. MS/MS data was used to search the mouse N-terminal database using the MASCOT search engine. The taxonomy was restricted to *Mus musculus;* fixed modifications: N-terminal acetylation and lysine acetylation; variable modification: oxidation of methionine; protease: Arg-C, missed cleavages: 1; peptide tolerance: 1.5Da, MS/MS tolerance: 0.6Da, instrument: ESI-TRAP, peptide charge: 1+, 2+ and 3+. Protein identifications with a mowse score greater than 20 were accepted as confident identifications. Peptide identifications corresponding to the ions observed in the MALDI-ToF spectrum are listed hers, for a complete list of identifications see supplementary data.

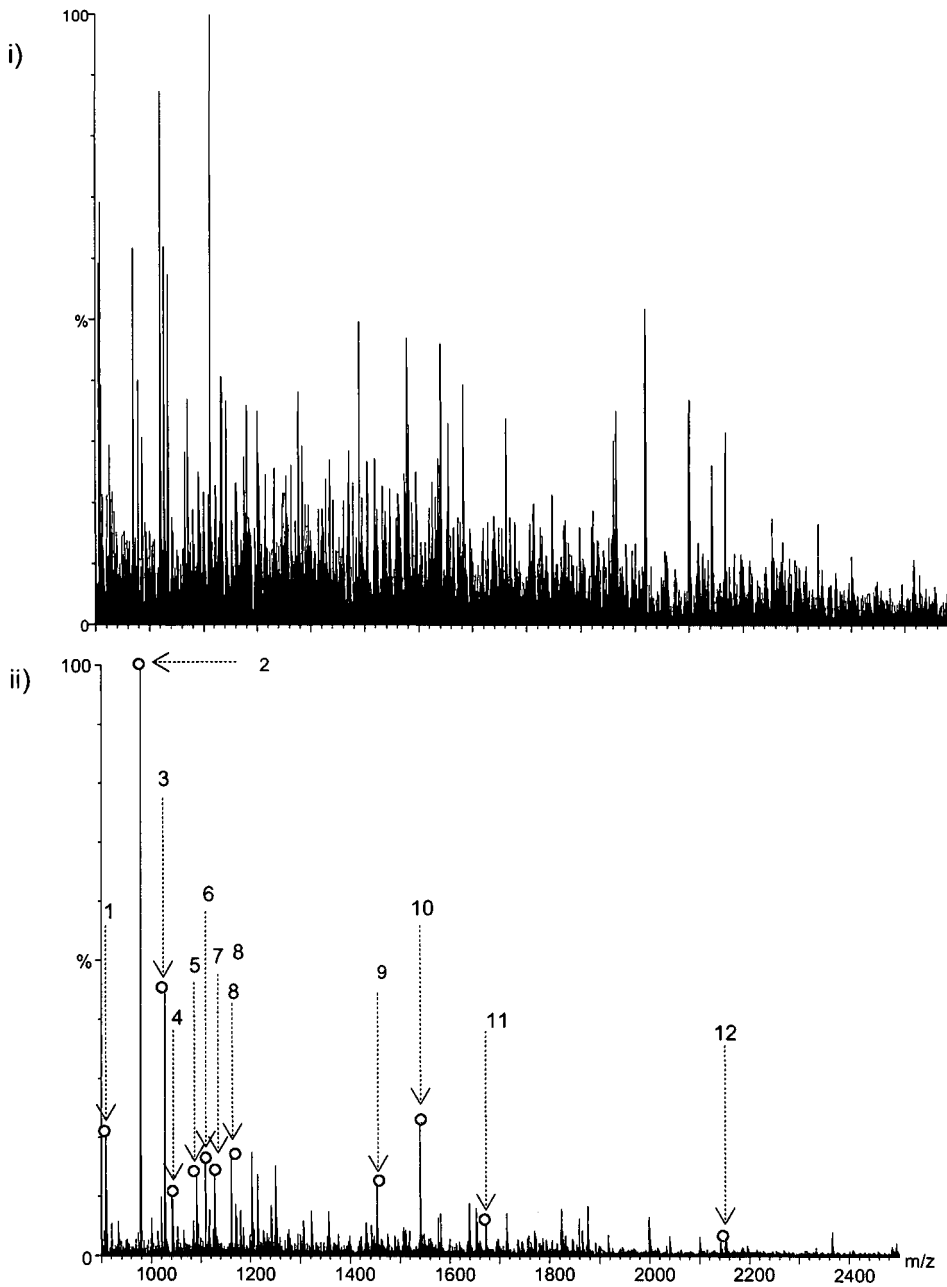


Figure 3.26. Isolated N-terminal peptides from *S. cerevisiae* cell lysate. The acetylated digest (a) and the N-terminal preparation (b) of *S. cerevisiae* proteins were analysed by MALDI-ToF MS using a laser energy of 30%All the major ions in the N-terminal spectra can be assigned to N-terminal peptides from *S. cerevisiae* (Table 3.8 represents the labelled ions).

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score
1	Phosphoglycerate mutase 1 (P00950)	907.59	PKLVLVR	M	400
2	Enolase 1 (P00924)	976.53	AVSKVYAR	M	98
3	Fructose-bisphosphate aldolase (P14540)	1025.59	GVEQILKR	M	180
4	40S ribosomal protein S3 (P05750)	1039.64	VALISKKR	M	36
5	60S ribosomal protein L1 (P53030)	1088.58	SKITSSQVR	M	84
6	Elongation factor 2 (P32324)	1107.54	VAFTVDQMR	M	54
7	40S ribosomal protein S15 (Q01855)	1126.61	SQAVNAKKR	M	57
8	60S ribosomal protein L37-A (P49166)	1159.60	GKGTPSFGKR	M	39
9	60S ribosomal protein L15-A (P05748)	1452.72	GAYKYLEELQR	M	87
10	Pyruvate decarboxylase (P06169)	1538.80	SEITLGKYLFER	M	189
11	60S ribosomal protein L5 (P26321)	1670.79	AFQKDAKSSAYSSR	M	69
12	Cytochrome C (P00044)	2150.41	TEFKAGSAKKGATLFKTR	M	87

Table 3.8. Identification N-terminal peptides from soluble *S. cerevisiae*, observed by MALDI-ToF MS.

The N-terminal peptide preparation of *S. cerevisiae* total cell lysate soluble proteins was analysed by LC-MS/MS using a three hour RP gradient. MS/MS data was used to search the mouse N-terminal database using the MASCOT search engine. The taxonomy was restricted to *S. cerevisiae;* fixed modifications: N-terminal acetylation and lysine acetylation; variable modification: oxidation of methionine; protease: Arg-C; missed cleavages: 1; peptide tolerance: 1.5Da; MS/MS tolerance: 0.6Da; instrument: ESI-TRAP; peptide charge: 1+, 2+ and 3+. Protein identifications with a Mowse score greater than 20 were accepted as confident identifications. Peptide identifications corresponding to the ions observed in the MALDI-ToF spectrum are represented, for a complete list of identifications see Supplementary data B.

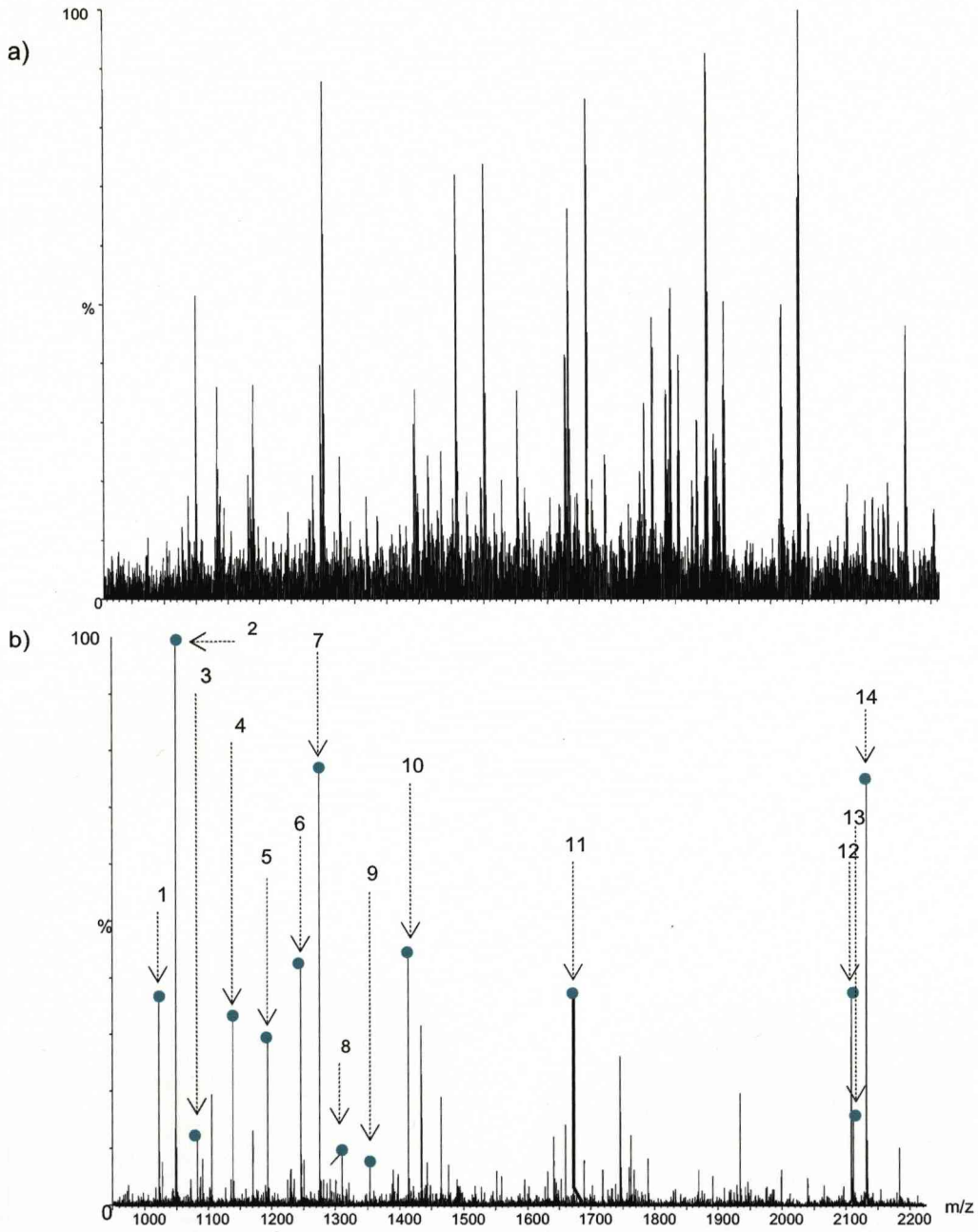


Figure 3.27. Isolated N-terminal peptides from *E. coli* cell lysate.
The acetylated digest (a) and the N-terminal preparation (b) of soluble *E. coli* proteins were analysed by MALDI-ToF MS using a laser energy of 30%. All the major ions in the N-terminal spectra can be assigned to N-terminal peptides from *E. coli* (Table 3.9 represents the labelled ions).

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score
1	UPF0304 protein yfbU (P0A8W8)	1021.43	MEMTNAQR		60
2	Elongation factor Tu (EF-Tu) (P0A6N1)	1048.52	SKEKFER	M	103
3	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase (P62707)	1081.69	AVTKLVLVR	M	46
4	Enolase (P0A6P9)	1138.71	SKIVKIIGR	M	113
5	10 kDa chaperonin (groES protein) (P0A6G1)	1192.61	MNIRPLHDR		39
6	Glyceraldehyde-3-phosphate dehydrogenase A (P0A9B4)	1244.69	TIKVGINGFGR	M	123
7	Copper resistance protein D (Q47455)	1274.67	MNDLIMIVIR		40
8	D,D-heptose 1,7-bisphosphate phosphatase (Q8FKZ1)	1299.72	AKSVPAIFLDR	M	42
9	Putative HTH-type transcriptional regulator yeaT (P76250)	1353.71	MNNLPLLNDLR	M	31
10	Elongation factor Ts (EF-Ts) (P0A6P1)	1412.79	AEITASLVKELR	M	285
11	Phosphoglycerate kinase (P0A799)	1671.89	SVIKMTDLDLAGKR	M	97
12	β-lactamase (P62593)	2107.12	pEQLIDWMEADKVAGPLLR	PG	84
13	Cysteine synthase A (P0ABK5)	2110.11	SKIFEDNSLTIGHTPLVR	M	64
14	β-lactamase (P62593)	2132.30	HPETLVKVKDAEDQLQR	SP	313

Table 3.9. Identification N-terminal peptides from soluble *E. coli*, observed by MALDI-ToF MS.

The N-terminal peptide preparation of *E. coli* total cell lysate was analysed by LC-MS/MS using a three hour RP gradient. MS/MS data was used to search the mouse N-terminal database using the MASCOT search engine. The taxonomy was restricted to *E. coli;* fixed modifications: N-terminal acetylation and lysine acetylation; variable modification: oxidation of methionine; protease: Arg-C; missed cleavages: 1; peptide tolerance: 1.5Da; MS/MS tolerance: 0.6Da; instrument: ESI-TRAP; peptide charge: 1+, 2+ and 3+. Protein identifications with a Mowse score greater than 20 were accepted as confident identifications. Peptide identifications corresponding to the ions observed in the MALDI-ToF spectrum are represented, for a complete list of identifications see Supplementary data B.

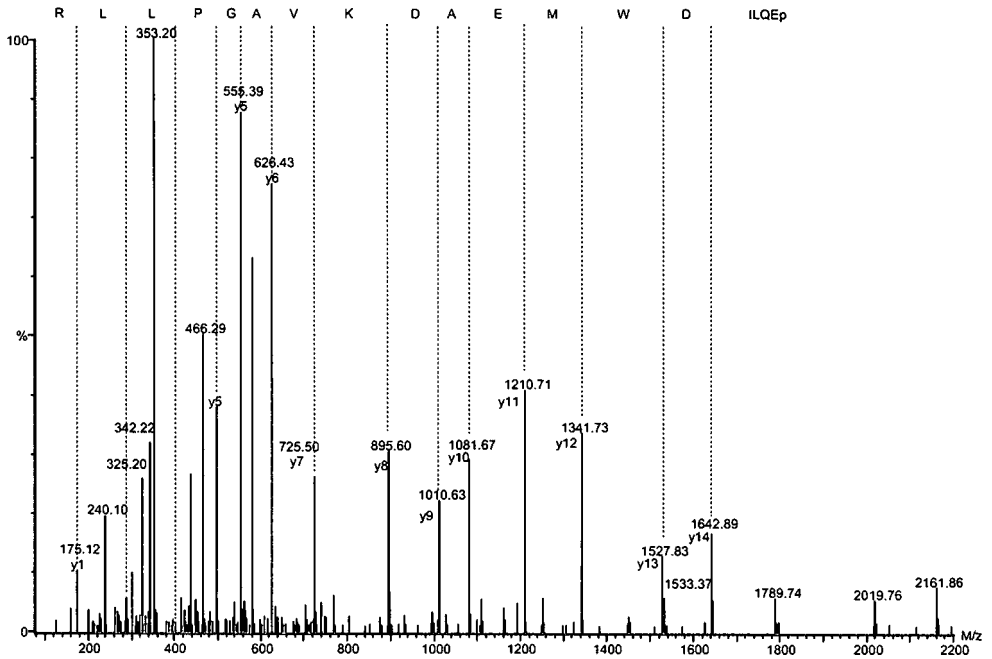


Figure 3.28. Sequencing of an unidentified internal peptide from the *E. coli* N-terminal preparation.

The *E. coli* N-terminal peptide preparation was desalted using a C18 ZipTip and introduced into the ESI-Q-ToF by direct infusion at a rate of 0.5µl/min. The precursor ion at $[M+2H]^{2+}$ 1055.02 was selected using the quadrupole and fragmented using a collision energy of 30%. The product ion spectra were processed using MaxENT3. The resulting MS/MS spectrum was sequenced *de novo* using the Biolynx software tool.

To screen for non N-terminal (internal or C-terminal) peptides, the MS/MS data was searched against Mascot, using the same parameters as in the N-terminal search. However, in this search N-terminal acetylation was chosen as a variable modification, to allow for any non N-acetylated peptides, in addition to N-acetylated, true N-terminal peptides. In all cases non N-terminal peptides were identified (Supplementary data C). In each case, the percentage of non N-terminal peptides represented <10% of the total identifications made. Moreover, many of the non N-terminal identifications made were matched with low confidence and are likely to be false positives.

The MALDI-ToF spectrum of the *E. coli* N-terminal preparation contained a major ion at m/z 2108.13, which was not identified by the N-terminal specific Mascot search. MS/MS analysis on the ESI-Q-ToF instrument, followed by manual *de novo* sequencing, revealed that this peptide is derived from the protein β-lactamase, with the sequence (R)pEQLIDWMEADKVAGPLLR (Figure 3.28). This internal peptide contained a modification at the N-terminus in which the glutamine residue was converted to pyroglutamate (pE). This spontaneous modification involves the cyclisation of N-terminal glutamine into pyroglutamate (Bateman *et al.,* 1990). When this reaction occurs the α-amino group becomes blocked and is consequently resistant to NHS-Sepharose coupling. As this peptide is arginine flanked, it is likely that this modification has occurred post-proteolysis (*in vitro)* and was not an *in vivo* modification.

A large percentage of the total protein identifications in *S. cerevisiae* were from ribosomal proteins (28%). These identifications were from both the 40S and the 60S subunits (24 and 26 proteins respectively).

From the lists of N-terminal peptide assignments, it is possible to explore the nature of the true N-terminus of the proteins identified. Figure 3.29 shows the frequency of amino acid residues at the N-terminal position in each of the three species analysed. In *E. coli* there is a large number of unprocessed N-terminal peptides (47% of total peptides identified) containing the N-terminal methionine residue. The percentage of unprocessed N-terminal peptides in the eukaryotic samples is much lower (15% in *S. cerevisiae* and 13% in mouse liver). SP cleavage products were observed in both *E. coli* and mouse liver samples. None of the proteins identified in *S. cerevisiae* showed evidence of SP cleavage, which is possibly due to the lack of secretary proteins present in the soluble preparation.

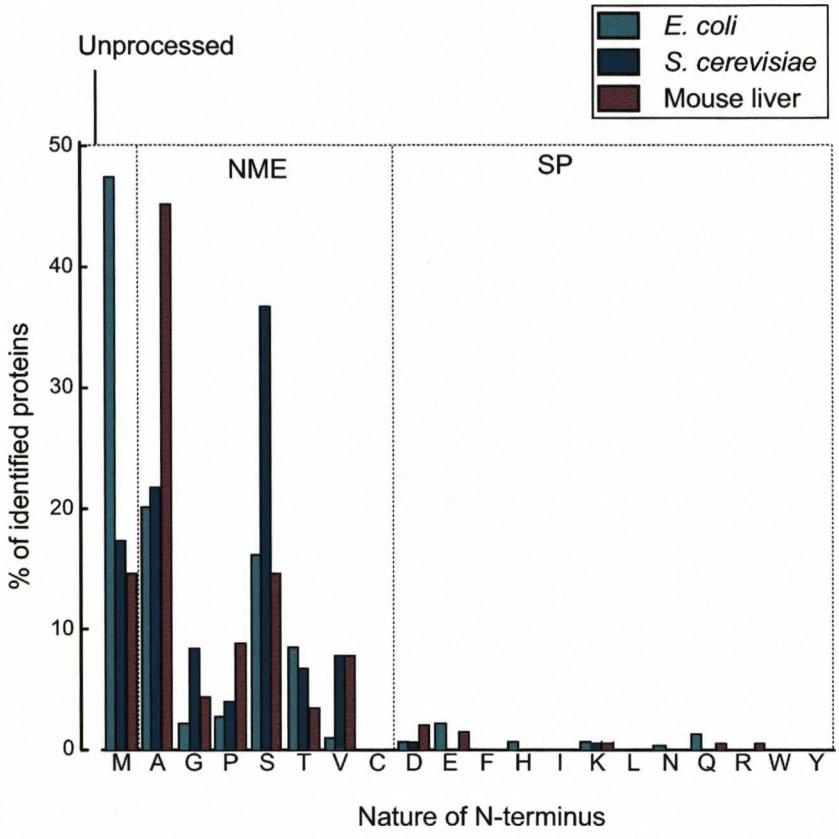


Figure 3.29. Nature of the N-terminal amino acid residues, determined by positional proteomics.
In total, 210 mouse liver, 185 *S. cerevisiae* and 333 *E. coli* proteins were identified by LC-MS/MS analysis on the ion trap instrument (data available in Supplementary data B). Data is from both blocked (N-acetylated) and unblocked N-termini and is grouped according to the N-terminal processing status.

3.9.11 Determination of naturally acetylated N-termini

It is well documented that between 80 and 90% of proteins from higher eukaryotes are naturally N-acetylated *in vivo* (Polevoda and Sherman, 2000). N^{α}-acetylation is much less common in prokaryotes, although it does still occur (Charbaut *et al.*, 2002).

The N-terminal methods described in this chapter both use acetylation to block N-termini. Therefore it is not possible to distinguish a naturally acetylated N-terminal from a chemically derivatised peptide (both produce a mass shift of +42Da). However, by switching to a labelled form of acetic anhydride it is possible to discriminate between chemical (*in vitro*) and natural (*in vivo*) N^{α}-acetylation. Deuterated acetic anhydride ($(C[^2H_3]CO)_2O$) reacts with primary amines, introducing a mass shift of +45Da instead of +42Da. This reagent was used to label N-terminal peptides from mouse liver, *S. cerevisiae* and *E. coli* soluble fractions. To illustrate the difference in mass between the incorporation of a standard (light) acetyl group and a deuterated (heavy) acetyl group, the standard peptide ACTH (Fragment 18-39, human) was acetylated by each reagent and analysed using MALDI-ToF MS (Figure 3.30). This peptide contains three possible acetylation sites. When acetylated, the pH of the environment was kept low (pH5), allowing partially acetylated forms to be analysed also.

Each of the three complex proteomes (mouse liver, *S. cerevisiae* and *E. coli*) were used to generate N-terminal preparations using $(C[^2H_3]CO)_2O$ as the amino group blocking reagent. The deuterated N-terminal preparations were subjected to analysis by MALDI-ToF MS and LC-MS/MS analysis. In order to discriminate between natural and chemical N^{α}-acetylation, the search parameters were adapted to incorporate a heavy acetyl group as a fixed modification on lysine residues and as a variable modification on the protein N-terminal. N^{α}-acetylation due to a normal acetyl group (+42Da) was also included as a variable modification. The search results were examined for the occurrence of either a light (+42Da), or a heavy (+45Da) acetyl group on the N-terminal. Data derived from the LC-MS/MS experiments is represented in Supplementary data B. In total, 98 mouse liver, 165 *S. cerevisiae* and 239 *E. coli* proteins were identified and their *in vivo* N-terminal acetylation status determined.

A comparison between the mouse liver N-terminal preparation derived using normal and deuterated acetic anhydride can be seen in Figure 3.31a. For each chemically added acetyl group, a mass shift of +3Da is observed. For example, the N-terminal of Sterol-4-alpha-carboxylate 3-dehydrogenase (Figure 3.31b; Q9R1J0) has an m/z value of 1370.78

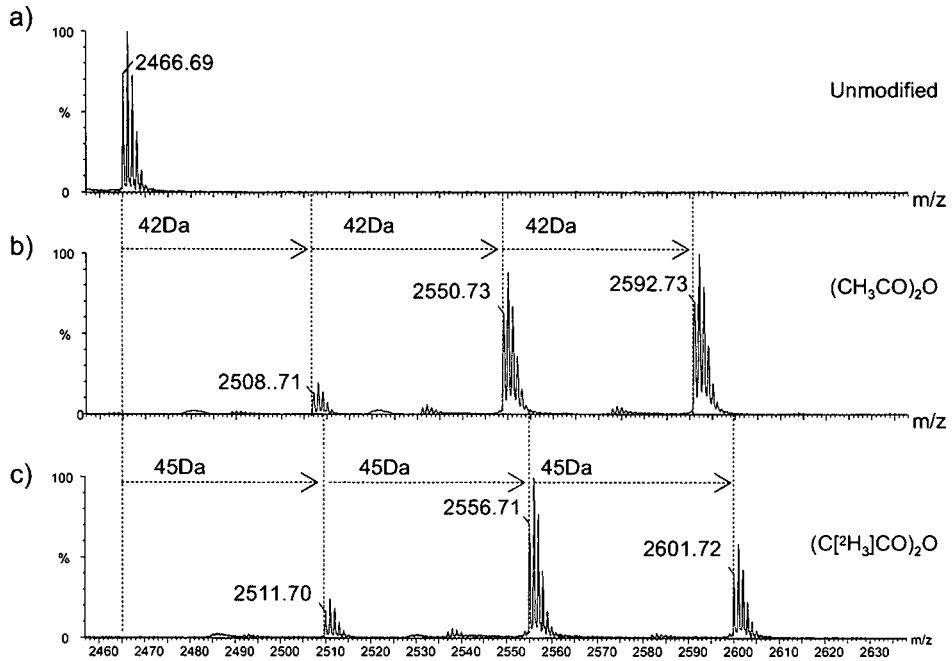
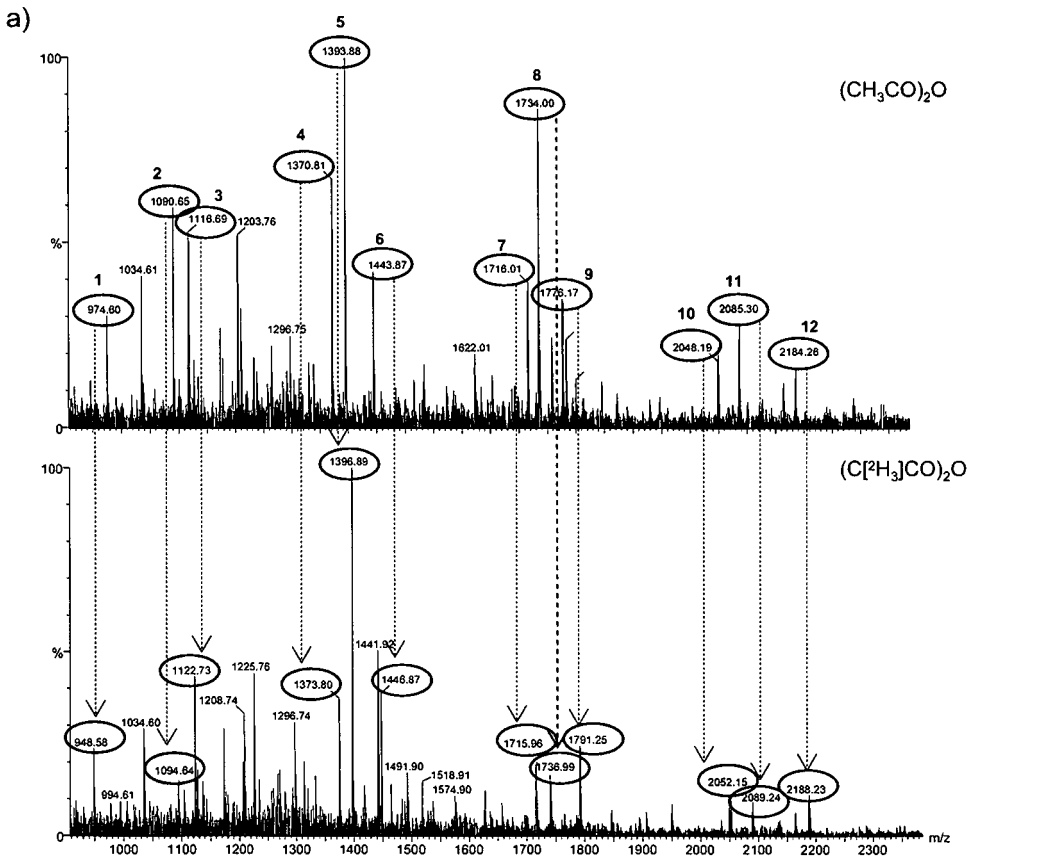


Figure 3.30. Acetylation of ACTH 18-39 using standard and deuterated acetic anhydride.

For comparative purposes, ACTH fragment 18-29 (Sigma) was analysed by MALDI-ToF MS in its unmodified form (a). The peptide (10µg) was acetylated with 1µl acetic anhydride (b) and 1µl deuterated acetic anhydride (c) for 1h at room temperature and analysed by MALDI-ToF MS. Acetylation reactions were performed in 20mM Na_2CO_3, pH8.5. Under these conditions the pH of the reaction drops substantially upon addition of acetic anhydride which favors incomplete reaction, thus ensuring the presence of multiply acetylated forms.

Figure 3.31. N-terminal purification of mouse liver soluble fraction using deuterated acetic anhydride.
Soluble proteins from mouse liver were subjected to two separate N-terminal isolation protocols, one using standard acetic anhydride, and the other using a stable isotope labeled variant $(C[^2H_3]CO)_2O$. The resulting preparations were analysed by MALDI-TOF MS (a). Mass shifts of +3Da were observed for the addition of each *in vitro* acetylated amino group. For non-lysine containing peptides a shift of +3Da is indicative of *in vitro* acetylation. For lysine containing peptides a mass shift of +3Da per lysine residue was observed, additional mass differences were evidence for *in vitro* N^{α}-acetylation. When no additional mass difference was observed *in vivo* acetylation was assumed. Sterol-4-alpha-carboxylate 3-dehydrogenase (b) and glutathione S-transferase P1 (c) are used as examples of *in vivo* and *in vitro* acetylation respectively. MALDI-ToF peptide ions are represented in Table 3.10.


a)
(CH3CO)2O
(C[2H3]CO)2O
m/z
b)
Sterol-4-alpha-carboxylate 3-dehydrogenase (Q9R1J0)
MEQAVHGESKR
3 Da
In vivo acetylation
b)
Glutathione S-transferase P 1 (P19157)
PPYTIVYFPVR
3 Da
In vitro acetylation

Spot	Protein (Ac)	Sequence	Mass $(CH_3CO)_2O$ (Da)	Mass (heavy)[b] $(C[^2H_3]CO)_2O$ (Da)	No. Lysine Residues	Δ Mass (Da)	Acetylated *in vivo*
1	60S ribosomal protein L39 (P62892)	SSHKTFR	945.47	948.47	1	3	Yes
2	SON protein (Q9QX47)	AADIEQVFR	1089.55	1092.55	0	3	No
3	Glyceraldehyde-3-phosphate dehydrogenase (P16858)	VKVGVNGFGR	1115.61	1121.61	1	6	No
4	Sterol-4-alpha-carboxylate 3-dehydrogenase (Q9R1J0)	MEQAVHGESKR	1370.64	1373.64	1	3	Yes
5	Glutathione S-transferase P 1 (P19157)	PPYTIVYFPVR	1392.74	1395.74	0	3	No
6	Glutathione S-transferase Yc (P30115)	AGKPVLHYFDGR	1442.73	1445.73	1	3	Yes
7	Liver fructose 1, 6 Bisphoshate (Q9QXD6)	ANHAPFETDISTLTR	1713.83	1713.83	0	6	No
8	Acyl-CoA-binding protein (P31786)	SQAEFDKAAEEKVR	1732.83	1735.84	1	6	No
9	Betaine--homocysteine S-methyltransferase (O35490)	APVAGKKAKKGILER	1775.03	1790.03	4	15	No
10	Peptidyl-prolyl cis-trans isomerase A (P17742)	VNPTVFFDITADDEPLGR	2046.99	2049.99	2	3	No
11	Sorbitol dehydrogenase (Q64442)	AAPAKGENLSLVVHGPGDIR	2084.10	2087.10	1	3	Yes
12	Peroxiredoxin 6 (O08709)	PGGLLLGDEAPNFEANTTIGR	2183.06	2186.06	0	3	No

Table 3.10. Determination of N^{α}-acetylation status of mouse liver N-terminal peptides.
The mass differences between $(CH_3CO)_2O$ (light) and $(C[^2H_3]CO)_2O$ (heavy) modified peptides were calculated and used to determine the true N^{α}-acetylation status of the protein. The full set of peptide identifications from this experiment are represented in Supplementary data B.

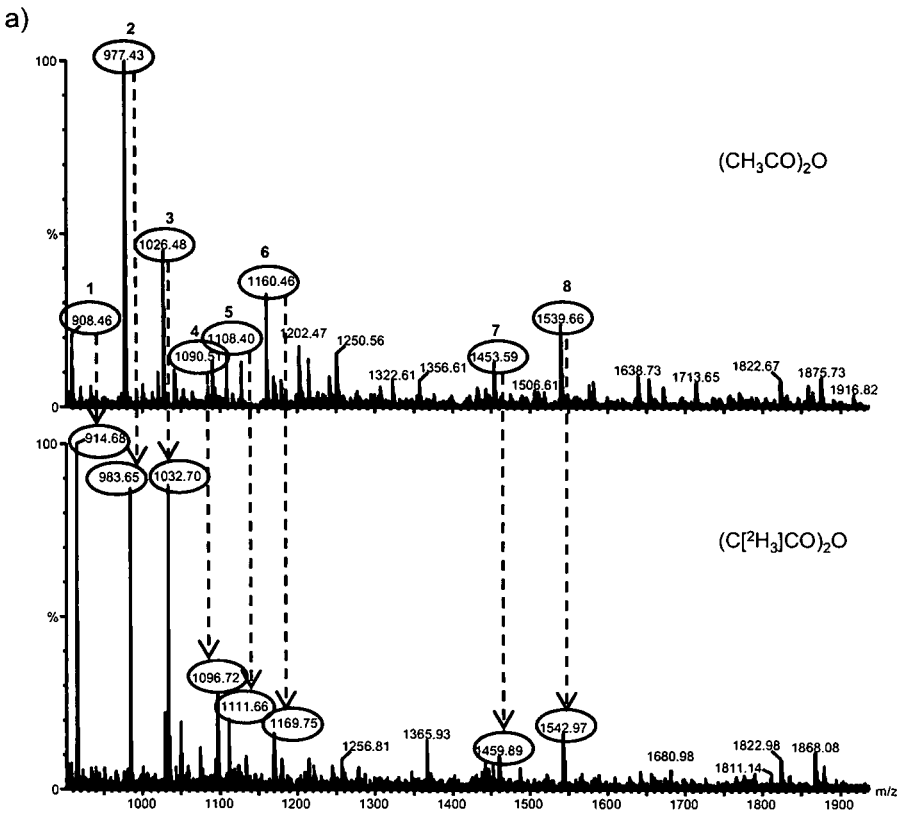
corresponding to the amino acid sequence (MEQAVHGESKR) with the addition of two acetyl groups. In the spectrum derived from the deuterated N-terminal preparation, this peptide is represented by the peak at 1373.78 m/z. The difference of 3Da between the two variants indicates that one chemically derived acetyl group has been added to the lysine residue, which in turn would indicate that this peptide is N^{α}-acetylated *in vivo.* In contrast, the N-terminal of glutathione S-transferase P1 (Figure 3.31c; P19157) has the m/z value of 1393.85, corresponding to the amino acid sequence (PPYTIVYFPVR) with the addition of one acetyl group. In the spectrum derived from the deuterated N-terminal preparation this peptide is represented by the peak at 1396.85 m/z. The mass difference of 3Da corresponds to the chemically induced acetyl group on the α-amino group, which implies that the N-terminal acetylation occurred *in vitro.* The data obtained from comparison of the two MALDI-ToF spectra is shown in Table 3.10. Naturally (*in vivo*) derived acetyl groups were observed in 87% (86 out of 98) of the proteins identified in mouse liver (Supplementary data B); a figure that is consistent with published data (Polevoda and Sherman, 2000).

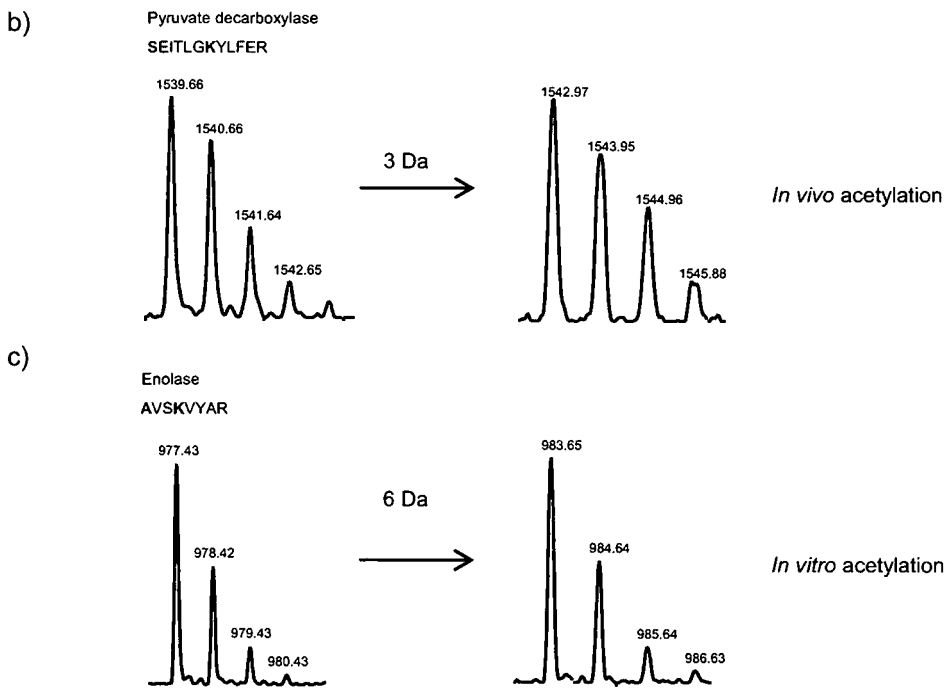
This analysis was repeated for *S. cerevisiae* (Figure 3.32; MALDI-ToF peak assignments are listed in Table 3.11). The data derived from the LC-MS/MS analysis is listed in Supplementary material B. Acetyl groups derived naturally (*in vivo*) were observed in 57 of the 128 proteins identified (44%), which is again consistent with published data (Lee *et al.,* 1989b).

The nature of the N^{α}-acetylated amino acid for both mouse liver proteins and *S. cerevisiae* proteins can be seen in Figure 3.33. N^{α}-acetylation occurs predominantly N-terminal alanine residues in mouse (53%) and serine residues in *S. cerevisiae* (80%). In both cases, the modification occurs on small amino acid residues that have undergone NME, in addition to a smaller level of N^{α}-acetylation on the uncleaved methionine residue (12% in mouse and 3% in *S. cerevisiae*). The lack of acetylation on proteins initiating with valine and glycine in *S. cerevisiae*, implies that acetylation on these residues plays some kind of evolutionary role, as these residues were acetylated in the mouse liver sample.

When performed on *S. cerevisiae*, this analysis provided data on the N^{α}-acetylation status of 54 ribosomal proteins. The results obtained in this positional proteomics N-terminal study were in total agreement to those obtained by Arnold *et al.* (Arnold *et al.,* 1999). In contrast to their time consuming and complex approach, the N-terminal positional proteomic strategy achieved the same goal but with substantially less experimental stages.

Figure 3.32. N-terminal purification of *S. cerevisiae* soluble proteins using deuterated acetic anhydride.
Soluble proteins from *S. cerevisiae* were subjected to two separate N-terminal isolation protocols, one using standard acetic anhydride, and the other using a stable isotope labeled variant $(C[^2H_3]CO)_2O$. The resulting preparations were analysed by MALDI-TOF MS (a). Mass shifts of +3Da were observed for the addition of each *in vitro* acetylated amino group. For non-lysine containing peptides a shift of +3Da is indicative of *in vitro* acetylation. For lysine containing peptides a mass shift of +3Da for each lysine residue was observed, additional mass differences were evidence for *in vitro* N^{α}-acetylation. When no additional mass difference was observed *in vivo* acetylation was assumed. Pyruvate decarboxylase (b) and enolase (c) are used as examples of *in vivo* and *in vitro* acetylation respectively. MALDI-ToF peptide ions are represented in Table 3.11.


a)
$(CH_3CO)_2O$
$(C[^2H_3]CO)_2O$
977.43
1026.48
1160.46
908.46
1108.40
1090.51
1202.47
1250.56
1322.61
1356.61
1453.59
1506.61
1539.66
1638.73
1713.65
1822.67
1875.73
1916.82
914.68
983.65
1032.70
1096.72
1111.66
1169.75
1256.81
1365.93
1459.89
1542.97
1680.98
1811.14
1822.98
1868.08
m/z


b)
Pyruvate decarboxylase
SEITLGKYLFER
1539.66
1540.66
1541.64
1542.65
3 Da
1542.97
1543.95
1544.96
1545.88
In vivo acetylation
c)
Enolase
AVSKVYAR
977.43
978.42
979.43
980.43
6 Da
983.65
984.64
985.64
986.63
In vitro acetylation

Spot	Protein (Ac)	Sequence	Mass $(CH_3CO)_2O$ (Da)	Mass $(C[^2H_3]CO)_2O$ (Da)	No. Lysine Residues	Δ Mass (Da)	Acetylated *in vivo*
1	Phosphoglycerate mutase 1 (P00950)	PKLVLVR	907.59	913.59	1	6	No
2	Enolase 1 (P00924)	AVSKVYAR	976.53	982.53	1	6	No
3	Fructose-bisphosphate aldolase (P14540)	GVEQILKR	1025.59	1031.59	1	6	No
4	60S ribosomal protein L1 (P53030)	SKITSSQVR	1088.58	1091.58	1	3	Yes
5	Elongation factor 2 (P32324)	VAFTVDQMR	1107.54	1110.54	0	3	No
6	60S ribosomal protein L37-A (P49166)	GKGTPSFGKR	1159.60	1168.60	2	9	No
7	60S ribosomal protein L15-A (P05748)	GAYKYLEELQR	1452.72	1458.72	1	6	No
8	Pyruvate decarboxylase (P06169)	SEITLGKYLFER	1538.80	1541.80	1	3	Yes

Table 3.11. Determination of N^{α}-acetylation status of *S. cerevisiae* N-terminal peptides.
The mass differences between $(CH_3CO)_2O$ (light) and $(C[^2H_3]CO)_2O$ (heavy) modified peptides were calculated and used to determine the true N^{α}-acetylation status of the protein. The full set of peptide identifications from this experiment are represented in Supplementary data B.

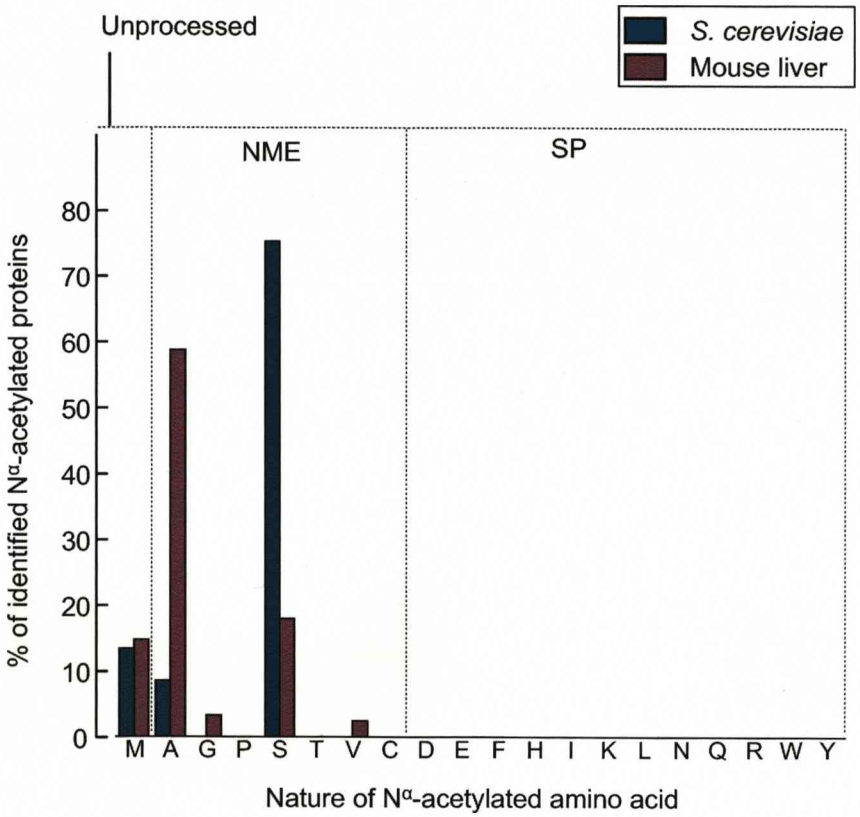


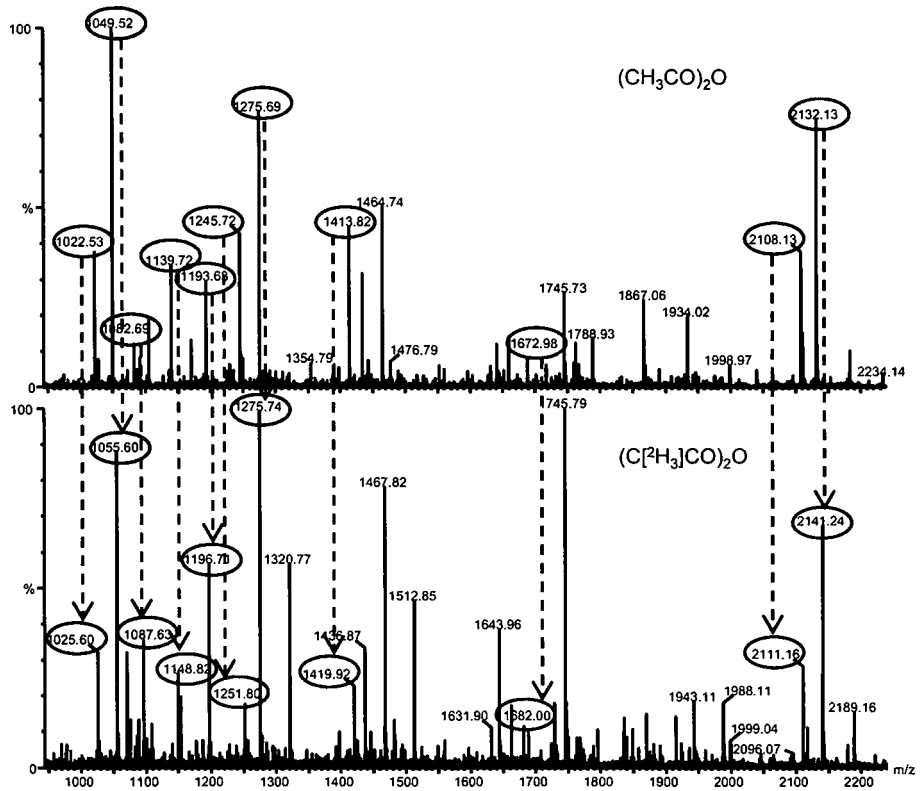
Figure 3.33. Nature of the Nα-acetylated N-terminal amino acid in mouse liver and *S. cerevisiae* proteins.
A total of 70 mouse liver proteins and 57 *S. cerevisiae* proteins were found to be Nα-acetylated *in vivo* (out of a total of 98 and 128 identified proteins respectively). Percentage values for the occurrence of each N-terminal amino acid were calculated. Data is grouped according to N-terminal processing status.

In contrast to eukaryotes, prokaryotes undergo a much lower rate of N^{α}-acetylation *in vivo*. To date, only five endogenous proteins have been reported to be N^{α}-acetylated in *E. coli* (Yoshikawa *et al.*, 1987; Tanaka *et al.*, 1989; Arai *et al.*, 1980; Smith *et al.*, 1996). To determine the extent to which N^{α}-acetylation occurs in the proteins contained within our *E. coli* N-terminal dataset, the N-terminal isolation procedure was performed on the soluble proteins from *E. coli*, using deuterated acetic anhydride as the acetylation reagent. The MALDI-ToF spectra obtained for N-terminal peptides acetylated using normal acetic anhydride and deuterated acetic anhydride are shown in Figure 3.34a. The majority of N-terminal peptide ions identified in this analysis were showed evidence of chemical acetylation. For example, the ion at m/z 2132.26 corresponding to β-lactamase (Figure 3.34c; HPETLVKVKDAEDQLQR) is represented in the deuterated form at m/z 2141.30. The +9Da difference is due to the addition of three chemical acetyl groups, corresponding to two lysine residues and the N-terminal amino group. This mass difference indicates that the N-terminal of enolase is acetyated *in vitro*. In contrast, the N-terminal of elongation factor Tu (EFTu) is represented in the light N-terminal spectra as m/z 1049.51 corresponding to the peptide sequence (Figure 3.34c; SKEKFER) plus three acetyl groups. In the spectra from the deuterated N-terminal peptides, this peptide is represented as 1055.60 m/z, giving a difference of +6Da corresponding to two acetyl groups (from two lysine residues), indicating that the protein is naturally N^{α}-acetylated. This is consistent with previously published data, as the N-terminal of EFTu is documented as being N^{α}-acetylated *in vivo* (Arai *et al.*, 1980). Both MALDI-ToF and LC-MS/MS data showed that the ion corresponding to the N-terminal peptide of copper resistance protein D (Figure 3.34d; MNDLIMIVIR) is N^{α}-acetylated *in vivo*. This previously unreported modification was validated by running a separate N-terminal isolation experiment in which the acetylation/blocking step was omitted. The rationale behind this experiment is that none lysine containing, *in vivo* N^{α}-acetylated peptides should be the only species resistant to NHS-Sepharose coupling. The resulting peptide preparation was analysed by MALDI-ToF MS and LC-MS/MS on the ion trap instrument (Figure 3.35). The base peak in this spectrum corresponds to the $[M+H]^+$ value of the N-terminal sequence for copper resistance protein D (verified by MS/MS spectrum; Figure 3.35c), thus proving that this peptide is N^{α}-acetylated *in vivo*. The ion at m/z 1745.55 was not fragmented sufficiently for interpretation and this peak remains uncharacterised.

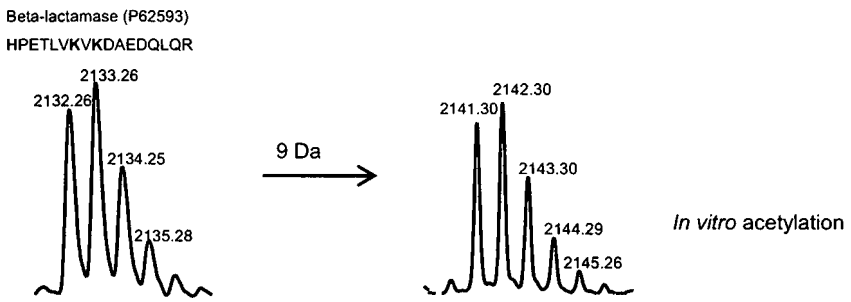
Other putative *E. coli* N^{α}-acetylated proteins identified in this study were hypothetical protein ydeR (P77294) and export protein SecB (P0AG86). However, it was not possible to obtain good quality MS/MS data from these two peptides by ESI-QToF or ESI-LTQ MS/MS.

Figure 3.34. N-terminal purification of *E. coli* soluble proteins using deuterated acetic anhydride.
Soluble proteins from *E. coli* were subjected to two separate N-terminal isolation protocols, one using standard acetic anhydride, and the other using a stable isotope labeled variant $(C[^2H_3]CO)_2O$. The resulting preparations were analysed by MALDI-TOF MS (a). Mass shifts of +3Da were observed for the addition of each *in vitro* acetylated amino group. For non-lysine containing peptides a shift of +3Da is indicative of *in vitro* acetylation. For lysine containing peptides a mass shift of +3Da for each lysine residue was observed, additional mass differences were evidence for *in vitro* Nα-acetylation. When no additional mass difference was observed *in vivo* acetylation was assumed. β-lactamase is used as an example of *in vitro* acetylation (b) and elongation factor Tu (c) and copper resistance protein D (d) are used as examples of *in vivo* acetylation MALDI-ToF peptide ions are represented in Table 3.12.

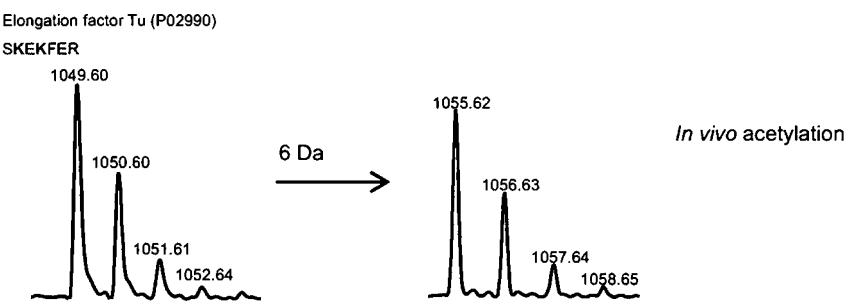
a)


(CH3CO)2O
(C[2H3]CO)2O
m/z

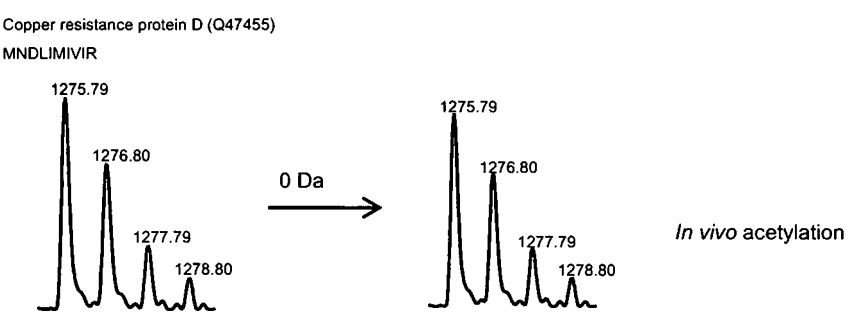
b)


Beta-lactamase (P62593)
HPETLVKVKDAEDQLQR
9 Da
In vitro acetylation

c)


Elongation factor Tu (P02990)
SKEKFER
6 Da
In vivo acetylation

d)


Copper resistance protein D (Q47455)
MNDLIMIVIR
0 Da
In vivo acetylation

Spot	Protein (Ac)	Sequence	Mass $(CH_3CO)_2O$ (Da)	Mass $(C[^2H_3]CO)_2O$ (Da)	No. Lysine Residues	Δ Mass (Da)	Acetylated *in vivo*
1	UPF0304 protein yfbU (P0A8W8)	MEMTNAQR	1021.43	1024.59	0	3	No
2	Elongation factor Tu (P02990)	SKEKFER	1048.52	1054.59	2	6	Yes
3	Phosphoglycerate mutase (P31217)	AVTKLVLVR	1081.69	1087.79	1	6	No
4	Enolase (P08324)	SKIVKIIGR	1138.71	1147.82	2	9	No
5	10 kDa chaperonin (groEL) (P0A6G1)	MNIRPLHDR	1192.61	1195.71	0	3	No
6	Glyceraldehyde 3-phosphate dehydrogenaseA (P06977)	TIKVGINGFGR	1244.6	1250.8	1	6	No
7	Copper resistance protein D (Q47455)	MNDLIMIVIR	1274.67	1274.67	0	0	Yes
8	Elongation factor Ts (EF-Ts) (P02997)	AEITASLVKELR	1412.79	1418.92	1	6	No
9	Phosphoglycerate kinase (P0A799)	SVIKMTDLDLAGKR	1671.89	1680.98	2	9	No
10	β-lactamase (P62593)	pEQLIDWMEADKVAGPLLR	2108.13	2111.16	1	3	No
11	β-lactamase (P62593)	HPETLVKVKDAEDQLQR	2132.3	2140.24	2	9	No

Table 3.12. Determination of N^{α}-acetylation status of *E. coli* N-terminal peptides.
The mass differences between $(CH_3CO)_2O$ (light) and $(C[^2H_3]CO)_2O$ (heavy) modified peptides were calculated and used to determine the true N^{α}-acetylation status of the protein. The full set of peptide identifications from this experiment are represented in Supplementary data B.

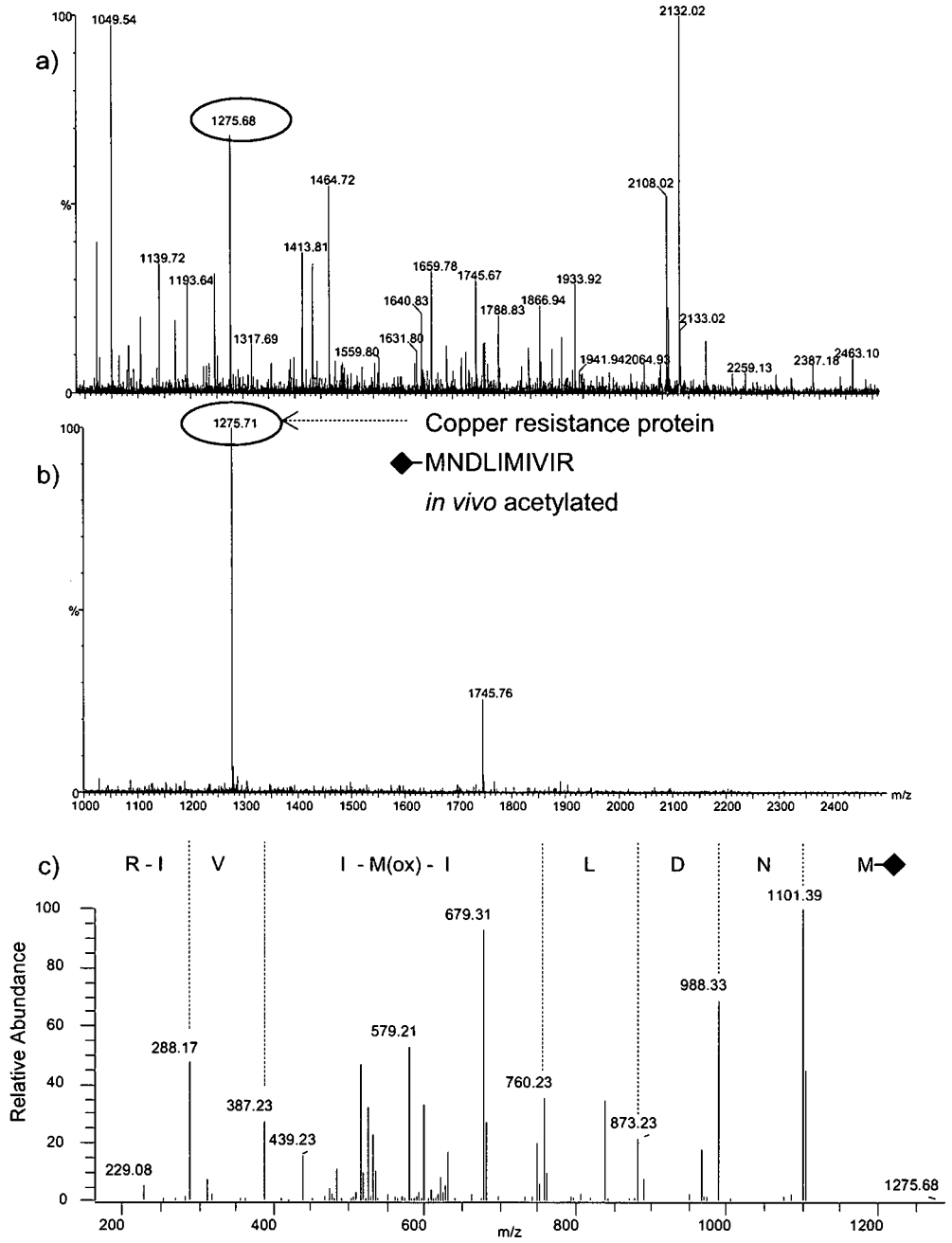


Figure 3.35. Isolation of naturally acetylated, non-lysine containing N-terminal peptides from *E. coli*.

E. coli soluble proteins (50µg) were digested with trypsin (1:50 enzyme substrate ratio). The resulting peptides were incubated with NHS-Sepharose (twice in total). The supernatant was removed by centrifugation and desalted using a C18 ZipTip prior to MALDI-ToF MS (b). For comparative purposes the previously characterised *E. coli* N-terminal preparation was also analysed (a). The peptides from the non-acetylated digest (b) correspond exclusively to naturally acetylated N-terminal peptides. The ion at m/z 1275.67 corresponds to copper resistance protein D, confirmed by LC-MS/MS analysis on the LTQ ion trap instrument (c). The ion at m/z 1745.71 remains unidentified.

3.10 Summary

The majority of proteins encoded in any genome undergo some form of N-terminal modification and as a result, N-terminal peptides are generally poorly represented in published datasets. Modification events (for example, NME, SP removal, N^{α}-acetylation and exopeptidase activity), hinder identification by routine proteomic approaches and make protein characterisation highly challenging (Meinnel and Giglione, 2008).

Positional proteomics strategies, that specifically target the N-terminal, provide a powerful tool to characterise the true N-termini of proteins. Unlike conventional methods based on Edman degradation, it is also possible to identify proteins that are naturally N^{α}-acetylated. The use of deuterated acetic anhydride ($(C[^2H_3]CO)_2O$) enables the true nature of N^{α}-acetylation to be determined, at the MS stage, either by comparison of MALDI-ToF spectra or by searching LC-MS/MS data for a combination of natural and chemical mass shifts. Application of this method to a prokaryote (where N-acetylation is rare) and eukaryotic samples (in which the modification is common) illustrates the broad applicability if the approach.

4. REDUCING THE COMPLEXITY OF HUMAN PLASMA USING POSITIONAL PROTEOMICS

Plasma is the liquid component of blood (comprising 55% of the total volume), in which the blood cells are suspended. It consists of about 90% water which provides the solvent for dissolving and transporting nutrients and gasses (principally oxygen, carbon dioxide and nitrogen). Approximately 8% of plasma consists of protein and the remaining 2% consists of amino acids, glucose and other nutrients (Table 4.1). Plasma also contains inorganic ions which are important in regulating cell function and maintaining homeostasis (reviewed in Van Wynsberghe, 1995; Table 4.2).

Serum is prepared from plasma by the removal of fibrinogen via its conversion to a fibrin clot, together with platelets. Varying amounts of other proteins are also removed in the fibrin clot, through both specific and non-specific interactions (Lundblad, 2005). However, despite conflicting statements in the literature (Adkins *et al.*, 2002), serum retains many coagulation factors such as factor IX, factor X and factor XI (Lundbald, 2005).

4.1 PLASMA PROTEINS

True or "classical" plasma proteins are those that carry out specific functions in the circulation. Proteins that, for example, serve as messengers between tissues (e.g. peptide hormones) or that leak into the blood from tissues are not classed as classical plasma proteins. True plasma proteins are largely secreted by the liver and intestine. These proteins perform a variety of roles including transportation of insoluble substances around the body, blood clotting, response to disease (inflammatory response) and protection from infection (immune response; Van Wynsberghe *et al.*, 1995).

As discussed in Chapter 1, the dynamic range of proteins in plasma spans ten orders of magnitude or more (Veenstra *et al.*, 2005), which is far greater than the measurement capability of current technologies. Furthermore, clinically relevant proteins that are present in the blood, for example, as a result of tissue damage, are masked by the overwhelming abundance of relatively few proteins. As much as 99% of the total protein mass in plasma is made up from 22 proteins (Tirumalai *et al.*, 2003; Figure 4.1).

Constituent	%
Water	90.0
Protein	8.0
Organic substances	1.1
Inorganic ions	0.9

Table 4.1. Constituents of human plasma as percentages of total volume.

Ion	Symbol	Concentration (mmol/l)
Sodium	Na^+	135-146
Potassium	K^+	3.1-5.2
Calcium	Ca^{2+}	2.1-2.7
Chloride	Cl^-	98-108
Hydrogen carbonate	HCO_3^-	23-31
Phosphate	PO_4^{2-}	0.7-1.4

Table 4.2. The normal range of concentration of inorganic ions in human plasma.

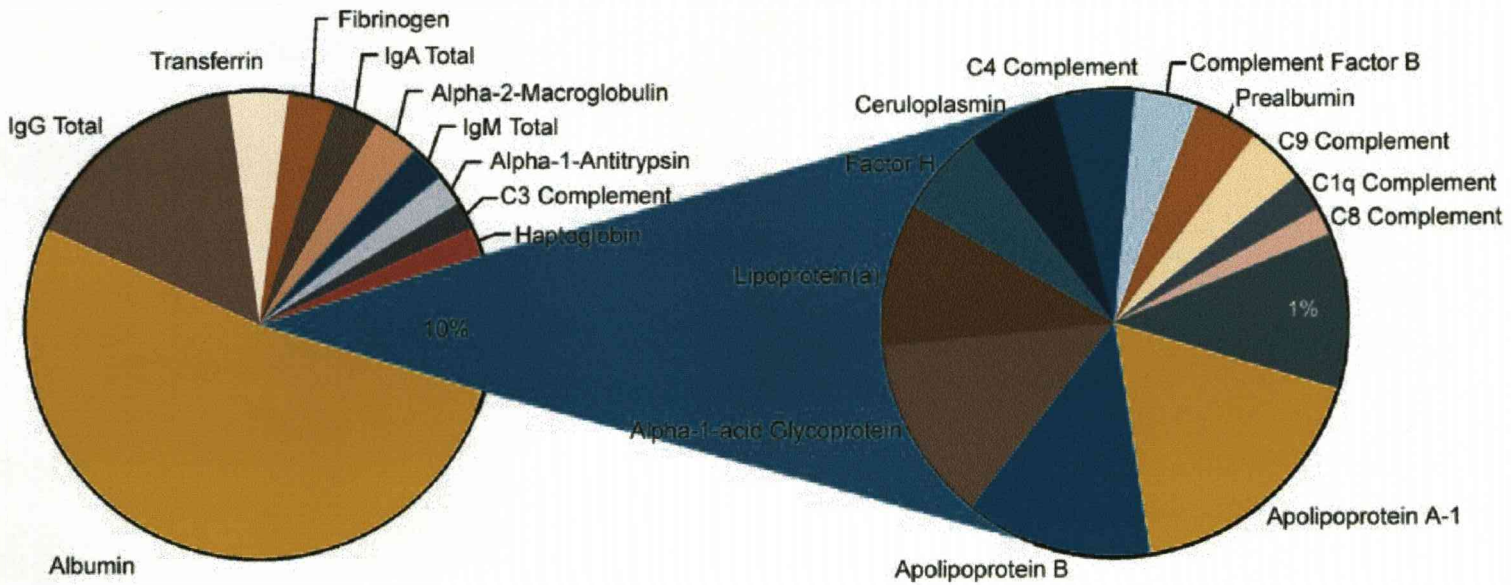


Figure 4.1. The relative abundances of proteins in plasma.
Twenty two proteins constitute ~99% of the protein content of human plasma (taken from Tirumalai *et. al.*, 2003).

Albumin is the most abundant protein in plasma and is responsible for around half of the total protein mass. A typical albumin molecule is present within the body for approximately 21 days and in its life span makes around 15,000 passes through the circulatory system, transporting different types of molecular cargo. The main function of albumin is to promote water retention in the blood, which in turn maintains osmotic pressure needed for efficient distribution of body fluids between intravascular compartments and body tissues (reviewed by Peters, 1995).

Albumin binds various proteins and peptides, for example, peptide hormones, interferon, serum amyloid A, glucagons, bradykinin and insulin (Baczynskyj *et al.*, 1994; Carter, 1981). These complexes form naturally occurring subproteomes (the albuminome). The association of albumin with small proteins poses a problem for simplification techniques based on albumin depletion, as the removal of large amounts of albumin will result in the subsequent co-depletion of these smaller, potentially relevant proteins that are specifically bound (Granger, 2005).

The plasma proteome is also dominated by the immunoglobulins, which represent around 15% of the protein content. Immunoglobulins, or antibodies, function in plasma to target and destroy foreign objects such as bacteria and viruses. Each immunoglobulin monomer consists of four chains, two heavy chains (~440 amino acids) and two light chains (~220 amino acids), held together by disulphide bonds (Figure 4.2). Each heavy and light chain is made up of two regions: a constant domain, which is identical in all immunoglobulins of a given class, and a variable domain, which resides at the tip of the immunoglobulin structure. It is this variable domain that is responsible for the immense variety of immunoglobulins that exist with slightly different tip structures. Each of these variants can bind to a different target, known as an antigen. This huge diversity of immunoglobulins allows the immune system to recognise an equally wide diversity of antigens. The unique part of the antigen recognised by the immunoglobulin is known as an epitope. These epitopes bind to their antigen in a highly specific interaction known as an induced fit, which allows the immunoglobulin to identify and bind to their antigen in the presence of millions of different protein molecules within the organism (Honjo and Habu, 1985; Davies *et al.*, 1990).

Immunoglobulins represent a unique class of proteins because of their complexity. There are thought to be in the order of 10 million different sequences of immunoglobulin in circulation in a normal adult (Anderson and Anderson, 2002). The basis for this diversity lies in the production of many different configurations of immunoglobulin in human lymphocytes. The

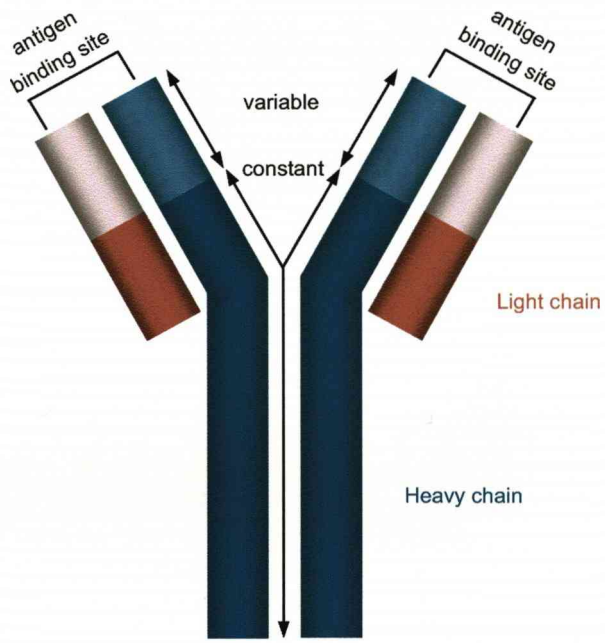


Figure 4.2. The immunoglobulin molecule.
Each immunoglobulin consists of four polypeptides: two heavy chains and two light chains joined to form a "Y" shaped molecule. The amino acid sequence in the tips of the "Y" varies greatly among different immunoglobulins. This variable region, composed of 110-130 amino acids, gives the immunoglobulin its specificity for antigen binding.

variable portions of an antibody's heavy and light chains is coded by several hundred coding regions and these regions, or gene segments, can be joined in different combinations to produce millions of immunoglobulins. There are five classes of immunoglobulin designated IgG, IgA, IgM, IgD and IgE. The constant region of the immunoglobulin heavy chain identifies its class and each class had a separate defensive role in the immune response (Van Wynsberghe *et al.*, 1995).

Antigen-antibody complexes trigger the mammalian complement system, which is a biological cascade that facilitates the removal of pathogens from an organism. The system consists of more than 30 proteins (Schmidt and Colten, 2000), which under normal circumstances are present in plasma as inactive zymogens. When stimulated by inflammation or an immune response, proteases in the system cleave specific components to release cytokines, initiating an amplification cascade of further cleavages. The main functions of the pathway include: host cell defense against microorganisms, elimination of immune complexes and apoptotic cells and the facilitation of adaptive immune responses. Complement is activated by three different pathways: classical, lectin and alternative. All three pathways share the common step of activation of complement component C3, but differ accordingly in the nature of regulation (reviewed by Reid and Porter, 1981).

Complement component C3 is regulated by conformational changes induced by selective proteolysis. The cleavage of mature C3 is mediated by enzyme complexes and generates the anaphylatoxin containing C3a and the major fragment C3b (Hugli, 1975). The C3a anaphylatoxin is generated when the C-terminal arginine is cleaved from fragment C3a and given the large amounts of carboxypeptidase present in plasma, is the major component of this fragment (de Bruijn and Fey, 1985). C3a anaphylatoxin is identical in terms of amino acid composition, molecular mass and N-terminal amino acid sequence as acylation-stimulating protein (ASP; Baldo *et al.*, 1993), which stimulates triacylglycerol synthesis in human adipocytes (Cianflone *et al.*, 1989) and is associated with obesity, cardiovascular disease, diabetes, and dyslipidemia (reviewed in Cianflone *et al.*, 2003). More recently, C3a anaphylatoxin was found in significantly higher levels in serum from patients with colorectal adenomas and carcinomas, than in healthy individuals. This suggests that a robust, reproducible clinical test for the quantification of C3a anaphylatoxin in plasma, could replace existing screening tests for colorectal cancer (Habermann *et al.*, 2006).

Plasma also contains a variety of proteins responsible for blood clot formation. The thrombin system consists of several blood proteins that become activated when a blood

vessel is traumatised and bleeding occurs. When platelets in the bloodstream come into contact with a damaged blood vessel, they release the enzyme thrombokinase, which brings about the conversion of the inactive zymogen prothrombin into the active thrombin which, in turn, catalyses the conversion of the soluble protein fibrinogen to the insoluble protein fibrin. This is a fibrillar protein that is polymerised to form a mesh that coagulates (in conjunction with platelets) to form a hemostatic plug or a clot. Calcium, vitamin K, and a variety of enzymes, known as clotting factors, are also necessary for efficient blood clotting (Doolittle, 1984).

4.2 Biomarkers

Proteins that normally function within tissues can leak into plasma as a result of cell death or tissue damage. These proteins include many of the most important diagnostic markers, also known as biomarkers. Examples of biomarkers include: cardiac troponins, creatine kinase, or myoglobin used in the diagnosis of myocardial infarction (McComb *et al.,* 1984).

A biomarker can be defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological response to a therapeutic intervention" (Atkinson *et al.,* 2001). The most valuable biomarkers are highly sensitive, specific and reproducible.

Despite the complexity issue discussed in Chapter 1, plasma remains the current sample of choice for proteomic biomarker discovery. In addition to containing proteins from every tissue and cell type within the body (Anderson *et al.,* 2004), plasma is easily accessible and can be repeatedly sampled from a patient. Therefore, plasma provides a valuable resource for monitoring protein expression changes and can provide important insights into disease status even before clinical symptoms are observed (Veenstra *et al.,* 2005). For diseases such as cancer early diagnosis is a crucial factor in the administration of effective treatment (Aebersold *et al.,* 2005). The ability to detect and characterise novel biomarkers associated with specific forms of cancer will therefore have great impact on public health. For this reason protein biomarker discovery in plasma has become one of the most significant applications of proteomics, with over 900 reported studies to date (PubMed). However, MS based protein identification approaches will only be successful if the protein of interest is reliably detectable using current instrumental capabilities.

In recent years, proteomics studies have generated a range of putative biomarkers (Polanski and Anderson, 2006). However, the list of candidates is limited by the general lack

of sensitivity and specificity exhibited by the majority of proteins found to date. This lack of sensitivity has discouraged many potential follow on studies from candidate biomarkers.

The lack of validation and verification of candidate biomarkers has led to a striking shortfall in the number of protein diagnostic tests emerging from proteomic studies. In order to bridge the gap between biomarker discovery and clinical use it will be necessary to develop innovative approaches to target known biomarkers and quantitatively detect them in multiple clinical samples.

The issue of biomarker validation also relates to preanalytical factors regarding the robustness of the clinical study. The problem of bias is described as "the most important threat to validity in biomarker research" (Ransohoff, 2005). Bias occurs if the affected and non-affected groups are handled in systematically different ways, introducing a diagnostic signature into one group but not the other. For example, if a sample of patients is much older than a sample of controls, then differences due to age may be misattributed to disease (Zhang *et al.,* 2005), or such differences may be induced through variations in the handling and processing of clinical specimens (Statland *et al.,* 1973; Villanueva *et al.,* 2006). Such factors may provide an explanation for "discrimination" in current proteomic studies (Baggerly *et al.,* 2004). The issue of bias reinforces the need for consistency in sample handling and the requirement for all samples in studies to be handled identically. The ability to provide consistent data from multiple samples is a critical factor in the application of proteomics to biomarker discovery and validation.

4.3 Strategies employed to study the human plasma proteome

As biomarkers are likely to occupy the low concentration range (present at ng/ml to pg/ml) of the plasma proteome, protein biomarker discovery from plasma presents an immense analytical challenge that requires a combination of experimental approaches and data handling. The PPP, described in Section 1.3, has integrated a collection of datasets produced by 55 laboratories worldwide; in order to generate a Core Dataset of 3020 confidently assigned plasma proteins (Omenn *et al.,* 2005). The PPP utilised a range of experimental protocols including depletion, fractionation, MS and immunoassay methods. In addition to the substantial dataset produced, the PPP has laid the groundwork for the development and validation of putative biomarkers linked with human disease.

An independent approach to the PPP, involved the extensive fractionation of human plasma using 2-D chromatography (105 SCX fractions, followed by RP-LC analysis) and MS/MS led to the identification of 2392 proteins with high confidence (Shen *et. al.*, 2005). Immunodepletion of high abundance proteins, followed by LC-MS/MS analysis, has previously been the method of choice for the preparation of complex biological samples (described in Section 1.5.3). However, these methods are of limited success as removal of several abundant species (as much as 90% of the overall protein amount), means that the dynamic range issue is not fully addressed. Furthermore, depletion of high-abundance proteins will lead to co-depletion, effectively removing a portion of the plasma proteome together with the abundant proteins (Granger, 2005).

Protein Equalizer™ technology (described in Section 1.5.5) has the ability to reduce dynamic range without depletion. This technology is based on the interaction of complex protein samples with a large, highly diverse, library of ligands coupled to poly(hydroxymethacrylate) beads. The library contains around 64 million different ligands which can, in principle, capture every protein species within a complex proteome (Righetti *et. al.*, 2006). This technology has been applied to the characterisation of the human plasma proteome (Sennels *et. al.*, 2007). In this study, separation of the ligand-bound (normalised) portion of the plasma proteome, using 1-D SDS-PAGE, followed by LC-MS/MS analysis of 20 gel slices (GeLC-MS/MS), led to identification of 1559 proteins with a 95% confidence rate.

One drawback to Protein Equalizer™ technology is linked to the nature of the 'normalisation' process, which enriches low abundance proteins whilst simultaneously diluting abundant proteins. This modification to protein abundance levels makes the process problematic for quantification studies, as information regarding the absolute amount of individual proteins in the starting material is lost. However, the technology provides a useful tool for the identification of low abundance proteins within complex proteomes, exhibiting high dynamic range.

4.4 Aims and objectives

Characterisation of the plasma proteome by standard bottom-up proteomic approaches leads to the generation of large amount of peptides from the most abundant proteins in the mixture (i.e., the classical plasma proteins). Signals from these peptides will consequently mask signals produced from the lower abundance portion of the plasma proteome. This chapter demonstrates the use of the N-terminal positional proteomics strategy to substantially reduce the complexity of a human plasma digest. Reducing the number of peptides representative of each protein in the mixture to, in theory, one peptide will improve the ability to identify proteins present in lower concentrations by minimising the suppression affect inflicted by peptides from abundant proteins.

Plasma proteins are subjected to a variety of proteases that target their N and C-termini (exopeptidases). As a result, proteins will potentially be present in a range of truncated forms. In addition to a reduction in peptide complexity, the N-terminal method will define the true N-terminal region of the proteins identified and can be used to identify substrates of exopeptidases activity.

The utilisation of a positional proteomics approach to prefactionation brings with it several improvements to the standard shotgun methodologies. Removal of internal peptides from a complex peptide mixture will result in a sample with substantially less proteolytic 'noise', which is a major problem in standard shotgun approaches. Furthermore, characterisation of the true N-termini of plasma proteins will serve to further improve our understanding of the plasma proteome.

The use of Protein Equalizer™ technology to reduce the dynamic range in human plasma is also demonstrated. This method effectively normalises the concentration range of complex protein mixtures without depletion of abundant proteins. When used in combination with the N-terminal enrichment protocol, this technology should serve to improve the number of plasma protein identifications, leading to the identification of candidate biomarkers which could then be quantified by alternative methods.

4.5 Results and discussion

Prior to N-terminal purification of human plasma proteins, the major protein species were separated using 1-D SDS-PAGE and identified by PMF. This analysis was performed to highlight the broad dynamic range exhibited by this complex biological sample.

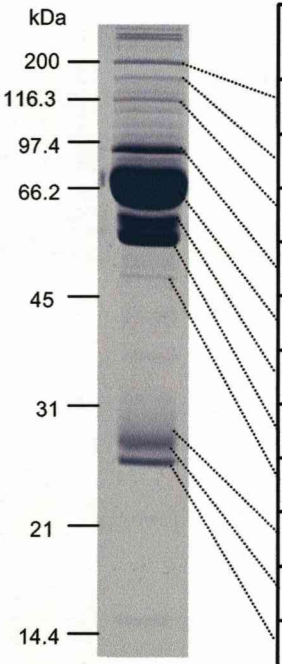
In-solution proteolysis of the entire plasma proteome was performed to demonstrate the suppression effect inflicted by serum albumin peptide ions, in terms of both MALDI-ToF MS and ESI-LC-MS/MS analysis using an extended three-hour gradient. This was necessary to appreciate the gain in coverage observed through targeted simplification of the sample by N-terminal abstraction.

Analysis of the N-terminal preparation of human plasma was performed on the LTQ ion trap instrument (in-house) and also on the LTQ Orbitrap (analysis performed by Gary Woffendin). The data was interrogated to assess the true nature of N-terminal peptides in human plasma, in addition to checking for the presence of 'leaked' internal peptides.

Human plasma was normalised using Protein Equalizer™ technology to reduce dynamic range. When a substantial reduction in concentration differences was achieved, compared with starting material (assessed using SDS-PAGE), the sample was subjected to LC-MS/MS analysis. The bound portion of the plasma proteome was also subjected to N-terminal purification to ascertain the effect of normalisation in combination with targeted peptide isolation in providing increased proteome coverage, compared with current strategies.

4.5.1 SDS-PAGE of human plasma proteins

To asses the complexity of human plasma, 15µg of protein was separated on a 1-D gel and proteins were visualised using Coomassie. The resulting 1-D profile is dominated by five major bands, corresponding to the most abundant proteins in the sample. A total of 11 protein bands were chosen for identification by PMF, including the five major bands and a further six weaker bands. To obtain identifications the bands were excised from the gel and subjected to in-gel proteolysis using trypsin. The digested peptide mixture was characterised by MALDI-ToF MS. Proteins were identified from their peptide mass fingerprint by manual searching against the SwissProt database using a locally implemented Mascot server. Search parameters allowed a single missed tryptic cleavage, carbamidomethyl modification of cysteine (fixed) oxidation of methionine (variable), and a peptide tolerance of ±150ppm. The taxonomic space was restricted to *Homo sapiens*. The proteins were all identified with high confidence (Table 4.3)



	Protein (Accession)	Mass (Da)	MOWSE score [a]	Coverage (%)	Peptides Matched
1	α-2-macroglobulin (P01023)	164.6	114 (57)	13	15
2	Ceruplasmin mature (P00450)	123.0	88 (54)	10	13
3	Complement factor b (P00751)	86.8	64(53)	10	6
4	Serotransferin (P02787)	79.3	256(55)	37	23
5	Serum albumin (P02768)	67.7	161 (54)	23	15
6	α -1-antitrypsin (P01009)	46.9	76 (55)	14	6
7	Ig gamma-1 chain C region (P01857)	36.6	60 (54)	18	6
8	Apolipoprotein A-IV (P06727)	45.34	100 (55)	29	11
9	Ig kappa chain V-III region SIE (P0183)	11.8	70 (54)	52	4
10	Apolipoprotein A-I (P02647)	30.8	123 (54)	33	10
11	Transthyretin (P05766)	16.0	45(55)	34	3

Table 4.3. Identification of plasma proteins using PMF.
Human plasma proteins (15µg, Sigma) were simplified by 1-D SDS-PAGE. Gel plugs from the major protein bands were excised and subjected to in-gel proteolysis with trypsin. The resulting peptide mixtures were analysed by MALDI-ToF MS and the monoisotopic masses were used to search the SwissProt database using the MASCOT search engine. The taxonomy was restricted to *Homo sapiens,* fixed modifications: carbamidomethylation of cysteine; variable modification: oxidation of methionine; protease: trypsin; missed cleavages: 1; peptide tolerance: 150pm. Significant matches above the given threshold (in brackets) were accepted as confident identifications.

as "classical" plasma proteins. Figure 4.3 shows the peptide coverage obtained for each identified protein. The MALDI-ToF spectra are shown in Supplementary material D.

4.5.2 In-solution tryptic digestion of human plasma proteins

Unlike the soluble cytosolic proteins found in skeletal muscle and liver, the majority of the proteins found in plasma are secreted. This implies that plasma proteins will have a greater abundance of disulphide bonds. For this reason, the solubilised human plasma (10µg, Sigma) was reduced and alkylated prior to proteolysis in order to remove the disulphide bonds between cysteine residues. The reduced and alkylated plasma was then TCA precipitated and ether washed, prior to overnight proteolysis with trypsin (50:1 substrate enzyme ratio). The resulting peptide mixture (1µl of digest diluted in 10µl 0.1% (v/v) TFA) was desalted using a C18 ZipTip and analysed by MALDI-ToF MS. Figure 4.4 shows that the entire digest of human plasma produces a complex MALDI-ToF spectrum. The *in silico* tryptic digestion of the human serum albumin (HSA) sequence revealed that most of the peaks in the spectrum can be attributed to this protein, shown using the peptide coverage map. Considering the high concentration of HSA in plasma it is not surprising that peptide ions originating from this protein dominate the mass spectrum.

Analysis by LC-MS/MS using the ion trap instrument led to the identification of only five abundant proteins from the sample (Table 4.4).

4.5.3 N-terminal isolation of human plasma proteins

Human plasma (50µg) was reduced and alkylated as before. The plasma was then TCA precipitated and ether washed to remove any residual reagents prior to acetylation. The dried pellet was resuspended in 20mM Na_2CO_3 (acetylation buffer) and acetylated using 1mg of sulfo NHS-acetate. The acetylated proteins were incubated with polymer bound Tris in order to remove excess reagent. The modified protein mixture was precipitated and washed (as before) prior to overnight proteolysis with trypsin (50:1 substrate enzyme ratio). A small amount of the peptide mixture (1µl of digest diluted in 10µl 0.1% (v/v) TFA) was desalted on a C18 ZipTip and subjected to MALDI-ToF analysis. The MALDI-ToF spectrum in Figure 4.5 corresponds to the acetylated digest of human plasma. In accordance with the non-acetylated tryptic digest (Figure 4.4), the majority of the high intensity signals can be attributed to albumin (Arg-C peptides). The peptide map indicates which peptides are present in the spectrum.

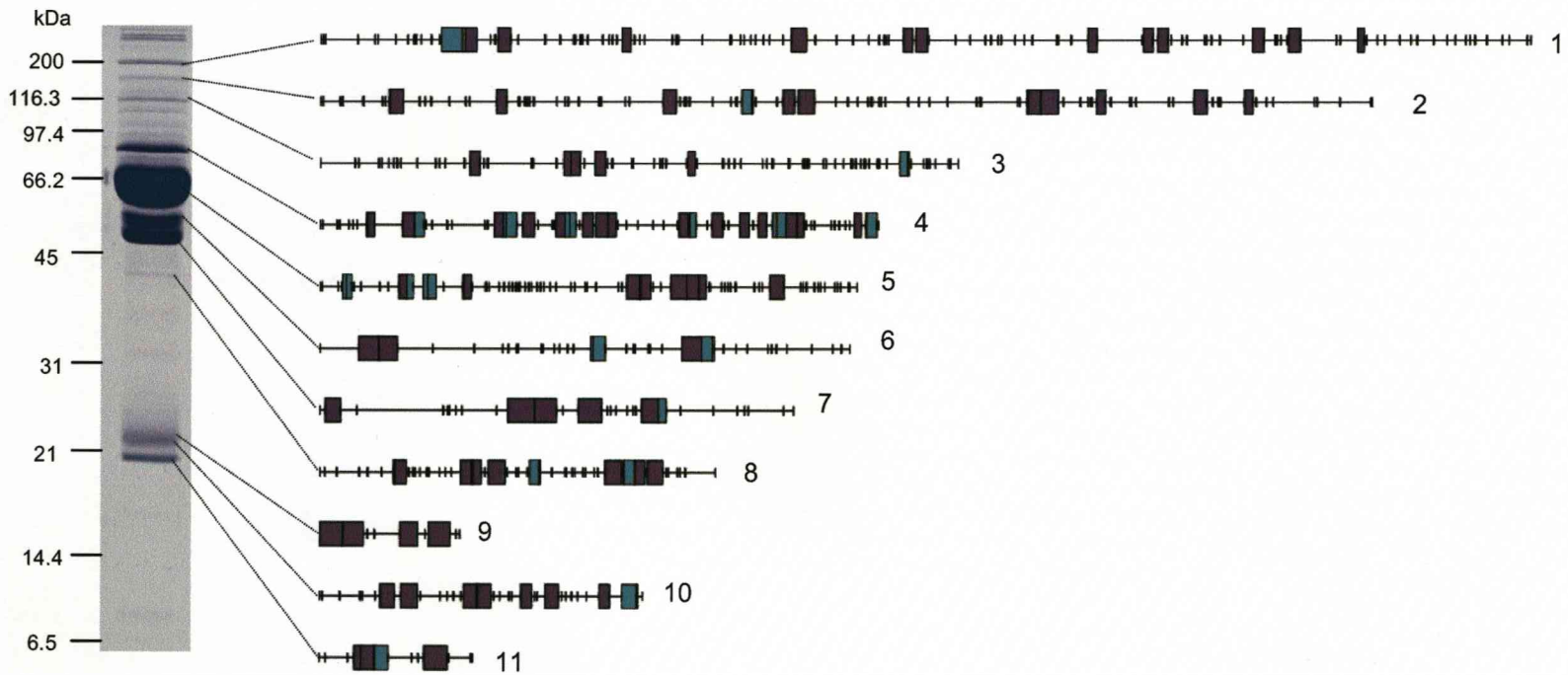


Figure 4.3. 1-D SDS-PAGE of human plasma proteins.
Peptide maps represent the coverage obtained from PMF analysis of human plasma proteins (Table 4.3). Limit tryptic peptides are represented in purple and peptides with one missed cleavage are in green.

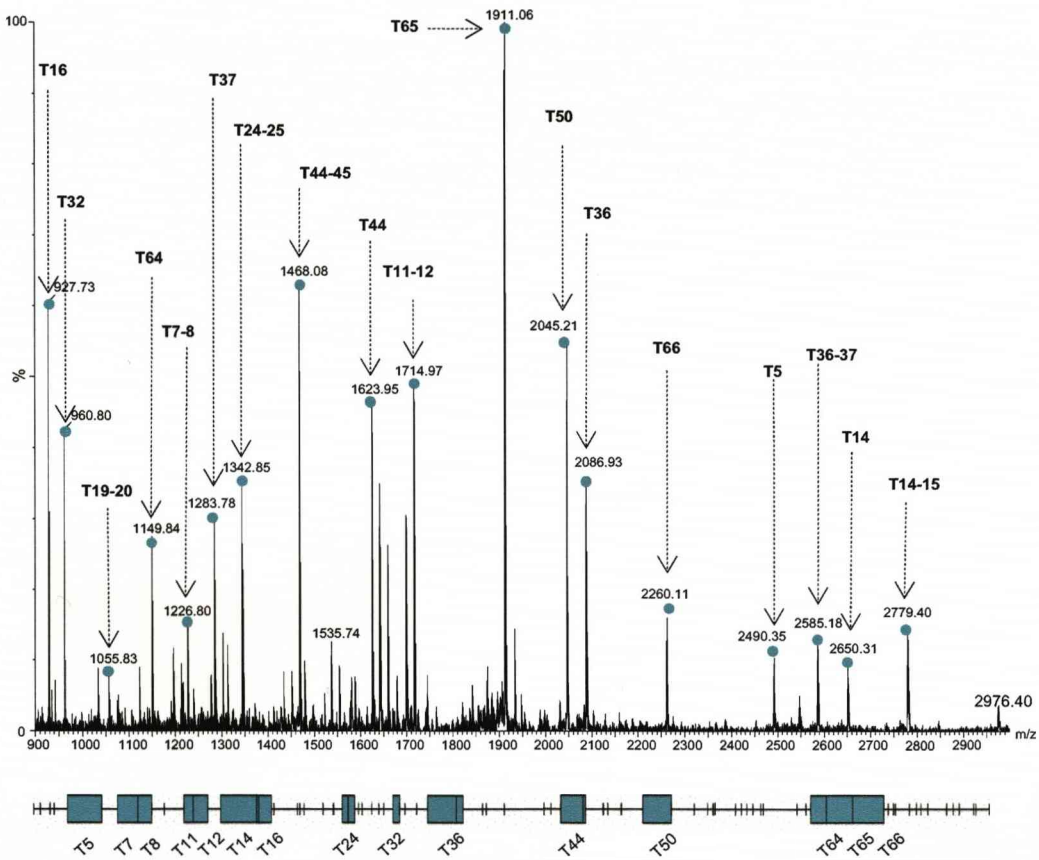


Figure 4.4. In-solution tryptic digest of human plasma.
Human plasma (10μg; SIGMA) was reduced and alkylated prior to digestion with trypsin (50:1 substate:enzyme). The resulting peptide mixture was desalted using a C18 ZipTip and analysed by MALDI-ToF MS. The majority of the major ions in the spectrum can be assigned to tryptic peptides from HSA. The peptide map represents coverage.

	Protein (Accession)	Mowse score	Coverage (%)	Peptides matched	Mass (kDa)
1	Serum albumin (P02768)	820	57	25	69.32
2	Serotransferrin (P02787)	216	18	8	79.28
3	Apolipoprotein A-I (P02647)	91	12	3	30.76
4	Ig kappa chain C region (P01834)	87	35	2	11.7
5	Ig lambda chain C region (P15814)	64	14	1	11.4

Table 4.4. Identification of human plasma proteins by in-solution tryptic digestion and LC-MS/MS analysis.
The peptide mixture derived from in-solution tryptic digestion of human plasma (10µg, Sigma) was analysed by LC-MS/MS using a three hour RP gradient. MS/MS data was used to search the SwissProt database using the MASCOT search engine. The taxonomy was restricted to *Homo sapiens;* fixed modifications: carbamidomethylation of cysteine; variable modification: oxidation of methionine; protease: trypsin; missed cleavages: 1; peptide tolerance: 1.5Da; MS/MS tolerance: 0.6Da; instrument: ESI-TRAP; peptide charge: 1+, 2+ and 3+. Protein identifications with a Mowse score greater than 50 were accepted as confident identifications.

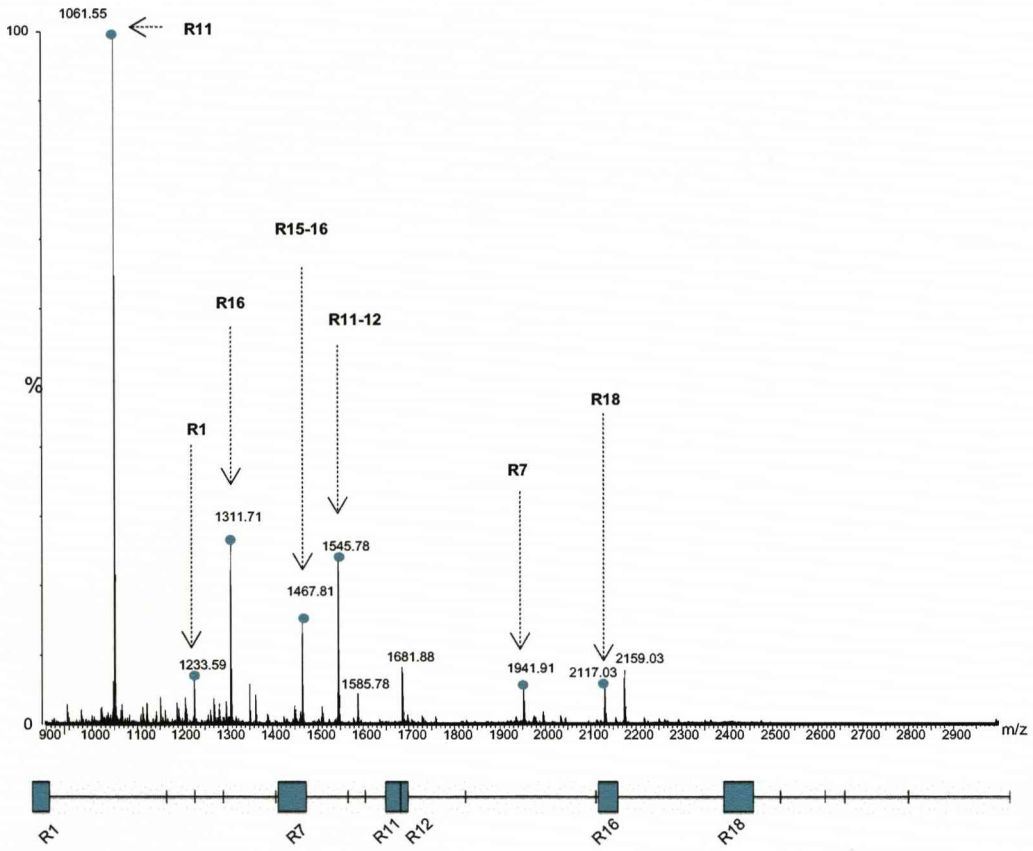


Figure 4.5. In-solution tryptic digest of acetylated human plasma.
Human plasma (50μg; SIGMA) was reduced, alkylated and acetylated prior to digestion with trypsin (50:1 substate:enzyme). The resulting peptide mixture was desalted using a C18 ZipTip and analysed by MALDI-ToF MS. The majority of the major ions in the spectrum can be attributed to arginine terminated tryptic peptides from HSA. Peptides containing internal lysines show an increased mass of +42Da for each residue. The peptide map represents coverage.

Following the N-terminal enrichment protocol, using NHS-activated Sepharose, the peptide mixture was substantially simplified (Figure 4.6a). The number of peptide ions arising from HSA was reduced from 18 to a single ion corresponding to the true N-terminal Arg-C peptide, produced as a result of pre-pro-peptide cleavage (Dugaiczyk *et al.*, 1982). A small peak corresponding to the O-acetylated form of the albumin N-terminal peptide was also present (Figure 4.7). NHS esters readily react with hydroxyl groups on serine groups located in linear His-X-Ser or Ser-X-His sequences (Miller, 1996), and although the N-terminal peptide preparation was treated with hydroxylamine (30 µmol), this was not sufficient to remove the high level of O-acetylated exhibited in this peptide.

Multiple LC-MS/MS analysis on the ion trap instrument using an extended (three hour) RP gradient led to the identification of 50 proteins by virtue of their N-terminal Arg-C peptide sequence (Table 4.5). A large percentage of the proteins identified were immunoglobulins (40%; Table 4.5b) and the remaining 60% were mainly classical plasma proteins (Table 4.5a). A small number of the proteins identified were non-classical plasma proteins including antigens and cell receptor proteins.

Although a greater number of protein identifications are obtained through application of the positional proteomics protocol, the data set was not as extensive as in samples analysed previously from mouse liver and *E. coli*. The failure to identify a higher number of low abundance proteins in the sample is probably a direct result of the large dynamic concentration range found in plasma. In human plasma as much as 99% of the total protein mass in plasma is made up from 22 proteins (Tirumalai *et al.*, 2003; Figure 4.1). Therefore, any additional proteins identified are from the remaining 1% of the proteome, making comprehensive proteome coverage highly challenging

The N-terminal preparation was also analysed by LC-MS/MS using the Orbitrap mass spectrometer, which can routinely deliver ~ ppm mass accuracy and resolution of 30,000-60,000. Under these conditions many N-terminal Arg-C peptides are distinct, in terms of mass, and can be identified by this parameter alone. To demonstrate this, extracted ion chromatograms for N-terminal Arg-C sequences from the most abundant proteins in plasma were prepared (Figure 4.8). In all cases the major ion in each MS spectrum, corresponded to the exact m/z value (to two decimal places) of the N-terminal peptide ion searched. Furthermore, the retention times of the isolated peptides were consistent with values predicted by the Sequence Specific Retention Calculator (SSRcalc) application (Krokhin *et al.*, 2004).

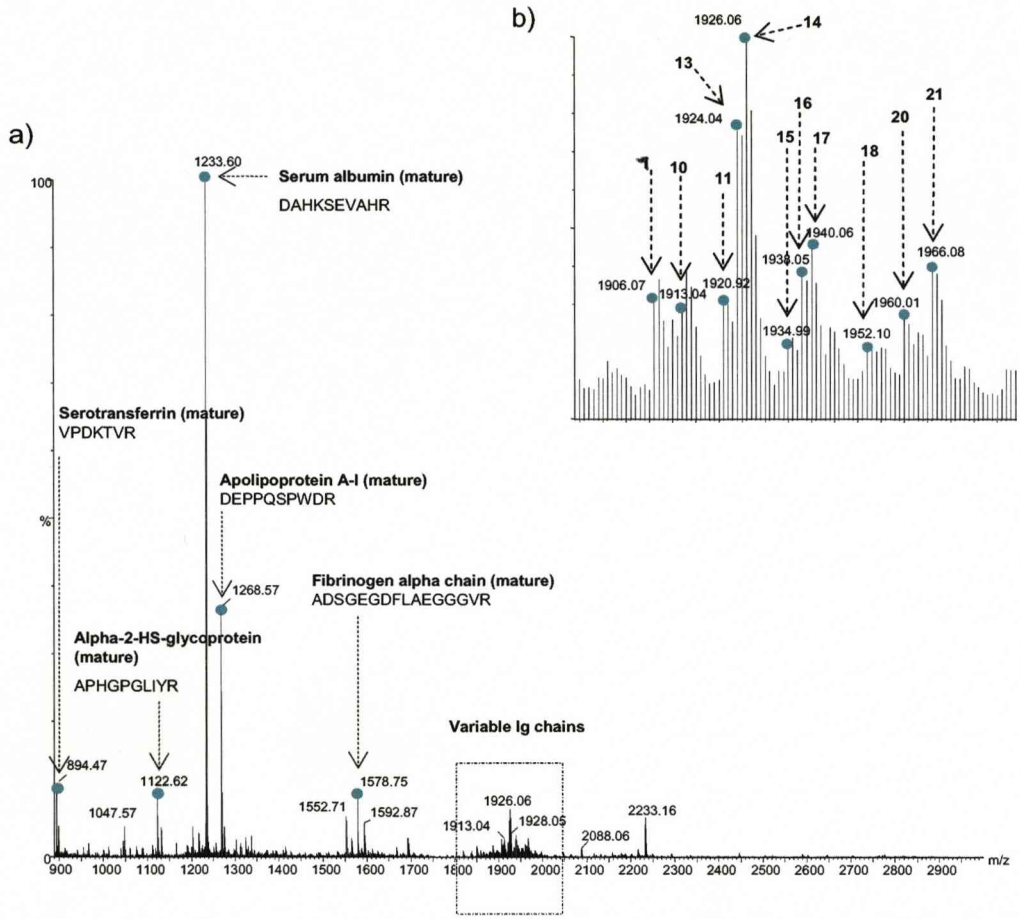


Figure 4.6. Human plasma N-terminal peptide preparation.
The acetylated tryptic peptides (50µg) were incubated with 1mg NHS-activated Sepharose. A small portion (10µl) of the of supernatant was removed and desalted using a C18 ZipTip and analysed by MALDI-ToF MS. The major ions in the spectrum (a) can be assigned to masses of acetylated, arginine terminated, mature N-terminal peptides originating from abundant plasma proteins. The cluster of ions around 1900Da (b) corresponds to N-terminal peptides from a varienty of Immunoglobulins. Peptide identifications were obtained by LC-MS/MS analysis on the ion trap instrument and are listed in Table 4.5.

a)

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score
1	Serotransferrin (P02787)	897.49	VPDKTVR	SP	85
2	Transmembrane protein 163 (Q8TC26)	1013.5	MEPAAGIQR		42
3	C3a anaphylatoxin (P01024)	1044.57	SVQLTEKR	M	35
4	Mitochondrial import receptor (O94826)	1081.5	GTAGPGTGGLPR		28
5	antigen MLAA-39 (Q96EA4)	1103.53	MEADITNLR		22
6	Alpha-2-HS-glycoprotein (P02765)	1121.6	APHGPGLIYR	SP	36
7	apolipoprotein A-IV (P06727)	1145.57	LEPYADQLR	SP	45
8	Glutathione S-transferase Mu 1 (P09488)	1175.60	PMILGYGDIR	M	41
9	Oxidoreductase HTATIP2 (Q9BUP3)	1200.63	AETEALSKLR	M	20
10	Serum albumin (P02768)	1232.6	DAHKSEVAHR	SP	201
11	Apolipoprotein A-I (P02647)	1267.55	DEPPQSPWDR	SP	185
12	Carboxypeptidase B2 (Q96IY4)	1327.73	FQSGQVLAALPR	SP	59
13	Antithrombin-III (P01008)	1520.74	HGSPVDICTAKPR	SP	48
14	Fibrinogen alpha chain (P02671)	1577.7	ADSGEGDFLAEGGGVR	SP	95
15	Serum amyloid P-component (P02743)	1585.83	HTDLSGKVFVFPR	SP	92
16	antigen NY-CO-43	1655.68	MEQGTSSSMTESSPR		32
17	Transcription factor ETV6 (P41212)	1658.76	SETPAQCSIKQER	M	21
18	Inter-alpha-trypsin inhibitor (Q14624)	1792.88	EKNGIDIYSLTVDSR	SP	36
19	FBXW12 mature	1826.88	SPTHQIQDPKHWNR	SP	25
20	Apolipoprotein E (P02649)	1836.91	KVEQAVETEPEPELR	SP	65
21	Tyrosin kinase receptor erbB-3 (P21860)	2104.86	EPCGGLCPKACEGTGSGSR	SP	29
22	cyclin G associated kinase	2116.1	SLLQSALDFLAGPGSLGGASGR	M	24
23	Alpha-1-antichymotrypsin (P01011)	2176.96	HPNSPLDEENLTQENQDR	SP	72
24	Guanylate kinase (Q16774)	2247.27	SGPRPVVLSGPSGAGKSTLLKR	M	47
25	Transthyretin (P02766)	2483.24	EAGPTGTGESKCPLMVKVLDAVR	SP	88
26	Apolipoprotein C-I (P02654)	2701.34	TPDVSSALDKLKEFGNTLEDKAR	SP	102
27	Replication protein A 30 kDa subunit	2776.19	SKSGFGSYGSISAADGASGGSDQLCER	M	22
28	HIV-1 like protein (O95081)	2997.44	VMAAKKGPGPGGGVSGGKAEAEAASE	M	26
29	Hemoglobin subunit beta (P68871)	3286.68	VHLTPEEKSAVTALWGKVNVDEVGGEA	M	77
30	Hemoglobin subunit alpha	3362.69	VLSPADKTNVKAAWGKVGAHAGEYGAE	M	95

Table 4.5. Identification of human plasma proteins by LC-MS/MS analysis of the N-terminal peptide preparation.
The N-terminal peptide preparation of human plasma was analysed by LC-MS/MS using a three hour RP gradient. MS/MS data was used to search the human N-terminal database using the MASCOT search engine. The taxonomy was restricted to *Homo sapiens;* fixed modifications: N-terminal acetylation and lysine acetylation; variable modification: oxidation of methionine; protease: Arg-C; missed cleavages: 1; peptide tolerance: 1.5Da; MS/MS tolerance: 0.6Da; instrument: ESI-TRAP; peptide charge: 1+, 2+ and 3+. Protein identifications with a Mowse score greater than 20 were accepted as confident identifications. Immunoglobulin N-terminal identifications are listed in (b), all other identifications are listed in (a).

b)

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score
1	immunoglobulin heavy chain variable region (ABM67294)	1140.55	GESGGGVVQPGR	M	36
2	immunoglobulin heavy chain variable region	1621.85	QVQLVQSGGGAVQPGR	M	54
3	Ig heavy chain V-III region CAM (P01768)	1651.85	QVELVESGGGVVQPGR	M	21
4	Ig heavy chain V-III region GA (P01769)	1665.83	QVELVESGGGAVQPGR	M	32
5	Ig heavy chain V-III region BRO (P01766)	1823.94	VQLEESGGGLVQPGGSLR	M	49
6	Ig kappa chain V-I region BAN (P04430)	1901.93	DIQLTQSPSSLSASVGDR	M	74
7	Ig kappa chain V-I region Mev (P01612)	1905.87	DVQMTQSPSSLSASVGDR	M	47
8	Ig kappa chain V-I region Wes (P01611)	1905.87	DIQMTQSPSSVSASVGDR	M	55
9	Ig heavy chain VHDJ region (BAC02290)	1908.99	EVQLVESGGGVVQPGGSL	M	102
10	Ig lambda chain V-III region SH (P01714)	1911.99	SELTQDPAVSVALGQTVR	M	66
11	Ig kappa chain V-I region DEE (P01597)	1919.89	DIQMTQSPSSLSASVGDR	M	28
12	Ig heavy chain V-III region LAY (P01775)	1921.03	AVQLLESGGGLVQPGGSLR	M	58
13	Ig heavy chain V-III region BRO (P01766)	1923.01	EVQLVESGGGLVQPGGSLR	M	47
14	Ig kappa chain V-III region SIE (P01620)	1925.01	EIVLTQSPGTLSLSPGER	M	105
15	Ig kappa chain V-I region CAR (P01596)	1933.91	DIQMTQSPSTLSASVGDR	M	98
16	Ig heavy chain V-III region TIL (P01765)	1937.02	EVQLLESGGGLVQPGGSLR	M	85
17	Ig kappa chain V-III region VG precursor (P04433)	1939.03	EIVLTQSPATLSLSPGER	M	62
18	Ig heavy chain V-III region JON (P01780)	1951.04	DVQLVESGGGLVKPGGSLR	M	25
19	Ig heavy chain V-III region BUT (P01767)	1951.04	EVQLVETGGGLIQPGGSLR	M	89
20	Ig kappa chain V-IV region Len (P01625)	1958.96	DIVMTQSPDSLAVSLGER	M	45
21	Ig heavy chain V-III region TRO (P01762)	1963.09	QVQLVQSGGGLVKPGGSL	M	88

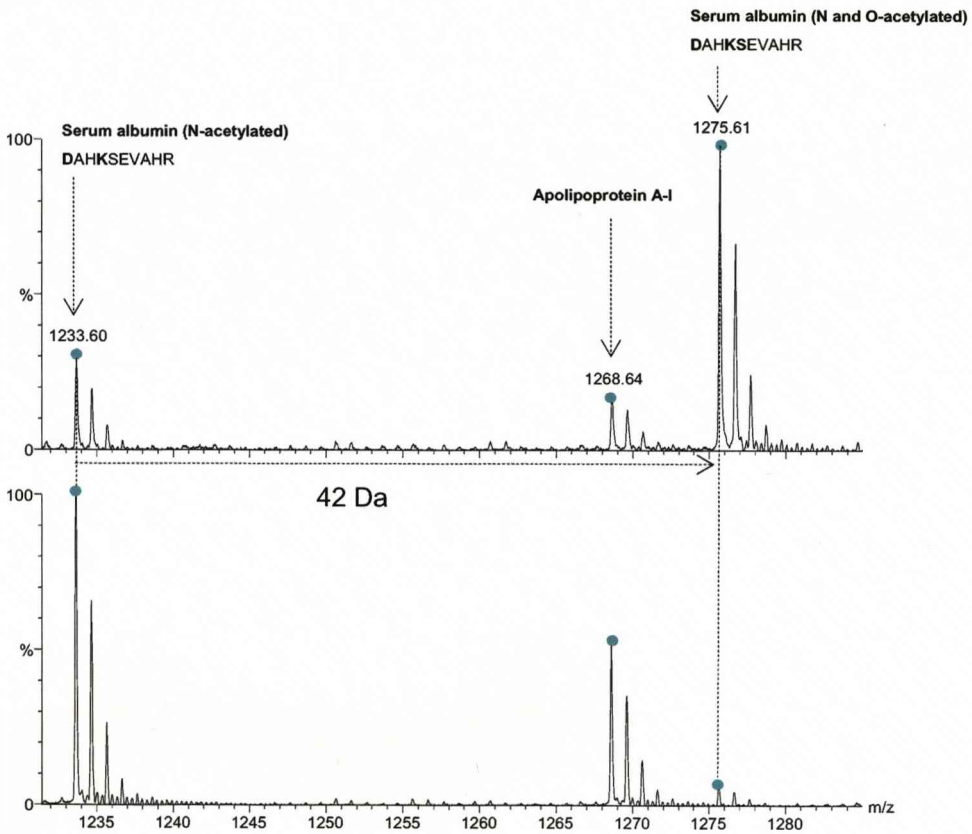


Figure 4.7. Effect of hydroxylamine treatment on the N-terminal of HSA .
N-terminal peptides recovered from NHS Sepharose incubation were treated with 1µl (30µmol) hydroxylamine in an attempt to reverse O-acetylation of hydroxyl groups. The upper spectrum includes two forms of the N-terminal HSA peptide, in which the major ion can be attributed to the O-acetylated form. The lower spectrum, obtained after treatment with hydroxylamine, shows incomplete removal of the O-acetylated form.

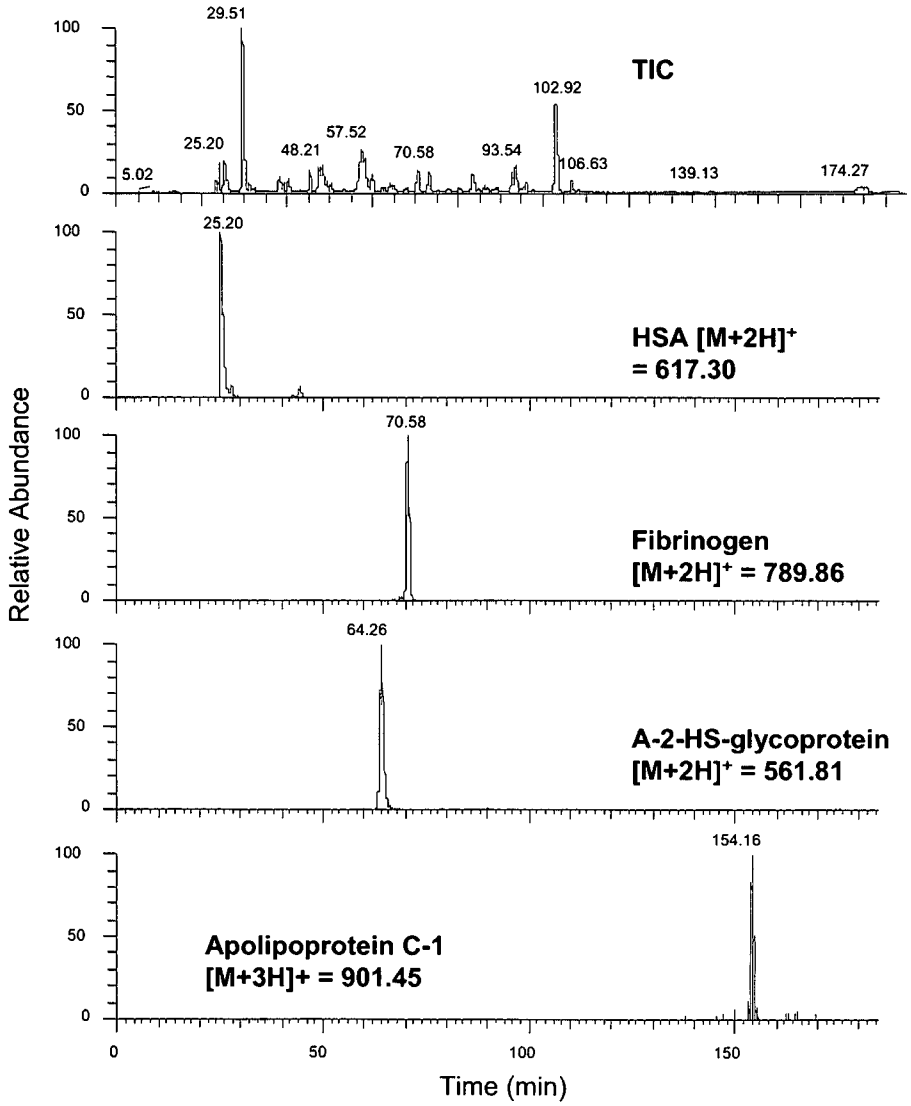


Figure 4.8. Chromatographic data from human plasma N-terminal peptides.
The N-terminal peptide preparation of human plasma was resolved by RP chromatography and analysed by high resolution ion trap MS on the Orbitrap instrument. The total ion chromatogram shows data for all peptides. Extracted ion chromatograms were prepared using the exact masses (+/- 0.01Da) for the true N-termini of four of the major proteins in human plasma.

4.5.4 Identification of multiple Immunoglobulin N-termini

A cluster of peaks visible in the 1900Da region of the MALDI-ToF mass spectrum (Figure 4.6b) were all identified, by LC-MS/MS on the ion trap instrument, as variable regions of unique Immunoglobulin N-terminal peptides (peak assignments shown in Table 4.5b). Variable regions of immunoglobulin molecules are contained within the amino (NH_2) terminal end of the polypeptide chain (amino acids 1-110; Davies *et al.,* 1990). When comparing one antibody to another, these amino acid sequences are quite distinct (Figure 4.2). The LC-MS/MS run identified in excess of 30 different immunoglobulin variable chains (Table 4.5b). The N-terminal immunoglobulin peptide sequences were aligned using ClustalW2 algorithm (Larkin *et. al.,* 2007; http://www.ebi.ac.uk/Tools/clustalw2/), in order to highlight the variation in amino acid sequence (Figure 4.9). The identification of multiple Immunoglobulin variable N-terminal peptides was seen in both pooled plasma samples (Sigma) and samples from individual subjects (data not shown).

In addition to its function in triggering an immune response, the immune system plays an important role in inhibiting cancer progression (Smyth *et al.,* 2001). Moreover, the immune system itself may act as a biomarker for the presence, type and possible even stage of cancer (Finn, 2005). It is possible that immunoglobulin profiling may provide insights into the molecular mechanisms of cancer progression (Tan, 2001). For this reason, utilisation of the N-terminal positional proteomics strategy could provide a rapid screening method for antibody profiling, by efficiently targeting the variable regions of immunoglobulin molecules.

4.5.5 Identification of N-termini derived from complement proteins

Activation of complement component C3 and subsequent generation of various fragments is fundamental to most of the biological activities exerted by complement (see Section 4.2). The sequence of complement C3 and the associated fragments can be seen in Figure 4.10. Because complement C3 is stored in SwisProt as a single protein entry (P01024; Figure 4.11), database searching using MS/MS data generated from the plasma N-terminal preparation, fails to identify N-terminal peptides from proteolytically derived complement fragments. In order to screen the N-terminal preparation for complement C3 fragments, extracted ion chromatograms were prepared manually on the accurate mass data from the Orbitrap mass spectrometer. The sequence of complement C3 precursor was retrieved using the Sequence Retrieval System (SRS; http://srs.ebi.ac.uk) by entering the SwissProt accession number. The N-terminal sequences for each of the fragments were taken from the feature table entry

```
P01768          QVELVESGGGV-VQPG---R
P01769          QVELVESGGGA-VQPG---R
P01767          QVQLVQSGGGA-VQPG---R
P01775          AVQLLESGGGL-VQPGGSLR
P01765          EVQLLESGGGL-VQPGGSLR
P01766          -VQLEESGGGL-VQPGGSLR
BAC02290        EVQLVESGGGV-VQPGGSLR
P01767          EVQLVETGGGL-IQPGGSLR
P01780          DVQLVESGGGL-VKPGGSLR
P01762          QVQLVQSGGGL-VKPGGSLR
P01611          DIQMTQSPSSVSASVGD--R
P01597          DIQMTQSPSSLSASVGD--R
P01596          DIQMTQSPSTLSASVGD--R
P01612          DVQMTQSPSSLSASVGD--R
P04430          DIQLTQSPSSLSASVGD--R
P01620          EIVLTQSPGTLSLSPGE--R
P04433          EIVLTQSPATLSLSPGE--R
P01625          DIVMTQSPDSLAVSLGE--R
P01714          -SELTQDPAVS-VALGQTVR
                   : :          *   *
```

Figure 4.9. Clustal alignment of immunoglobulin N-terminal (Arg-C) peptides.
Sequences were aligned using the ClustalW2 algorithm.

MGPTSGPSLL LLLLTHLPLA LGSPMYSIIT PNILRLESEE TMVLEAHDAQ GDVPVTVTVH
DFPGKKLVLS SEKTVLTPAT NHMGNVTFTI PANREFKSEK GRNKFVTVQA TFGTQVVEKV
VLVSLQSGYL FIQTDKTIYT PGSTVLYRIF TVNHKLLPVG RTVMVNIENP EGIPVKQDSL
SSQNQLGVLP LSWDIPELVN MGQWKIRAYY ENSPQQVFST EFEVKEYVLP SFEVIVEPTE
KFYYIYNEKG LEVTITARFL YGKKVEGTAF VIFGIQDGEQ RISLPESLKR IPIEDGSGEV
VLSRKVLLDG VQNPRAEDLV GKSLYVSATV ILHSGSDMVQ AERSGIPIVT SPYQIHFTKT
PKYFKPGMPF DLMVFVTNPD GSPAYRVPVA VQGEDTVQSL TQGDGVAKLS INTHPSQKPL
SITVRTKKQE LSEAEQATRT MQALPYSTVG NSNNYLHLSV LRTELRPGET LNVNFLLRMD
RAHEAKIRYY TYLIMNKGRL LKAGRQVREP GQDLVVLPLS ITTDFIPSFR LVAYYTLIGA
SGQREVVADS VWVDVKDSCV GSLVVKSGQS EDRQPVPGQQ MTLKIEGDHG ARVVLVAVDK
GVFVLNKKNK LTQSKIWDVV EKADIGCTPG SGKDYAGVFS DAGLTFTSSS GQQTAQRAEL
QCPQPAARRR RSVQLTEKRM DKVGKYPKEL RKCCEDGMRE NPMRFSCQRR TRFISLGEAC
KKVFLDCCNY ITELRRQHAR ASHLGLARSN LDEDIIAEEN IVSRSEFPES WLWNVEDLKE
PPKNGISTKL MNIFLKDSIT TWEILAVSMS DKKGICVADP FEVTVMQDFF IDLRLPYSVV
RNEQVEIRAV LYNYRQNQEL KVRVELLHNP AFCSLATTKR RHQQTVTIPP KSSLSVPYVI
VPLKTGLQEV EVKAAVYHHF ISDGVRKSLK VVPEGIRMNK TVAVRTLDPE RLGREGVQKE
DIPPADLSDQ VPDTESETRI LLQGTPVAQM TEDAVDAERL KHLIVTPSGC GEQNMIGMTP
TVIAVHYLDE TEQWEKFGLE KRQGALELIK KGYTQQLAFR QPSSAFAAFV KRAPSTWLTA
YVVKVFSLAV NLIAIDSQVL CGAVKWLILE KQKPDGVFQE DAPVIHQEMI GGLRNNNEKD
MALTAFVLIS LQEAKDICEE QVNSLPGSIT KAGDFLEANY MNLQRSYTVA IAGYALAQMG
RLKGPLLNKF LTTAKDKNRW EDPGKQLYNV EATSYALLAL LQLKDFDFVP PVVRWLNEQR
YYGGGYGSTQ ATFMVFQALA QYQKDAPDHQ ELNLDVSLQL PSRSSKITHR IHWESASLLR
SEETKENEGF TVTAEGKGQG TLSVVTMYHA KAKDQLTCNK FDLKVTIKPA PETEKRPQDA
KNTMILEICT RYRGDQDATM SILDISMMTG FAPDTDDLKQ LANGVDRYIS KYELDKAFSD
RNTLIIYLDK VSHSEDDCLA FKVHQYFNVE LIQPGAVKVY AYYNLEESCT RFYHPEKEDG
KLNKLCRDEL CRCAEENCFI QKSDDKVTLE ERLDKACEPG VDYVYKTRLV KVQLSNDFDE
YIMAIEQTIK SGSDEVQVGQ QRTFISPIKC REALKLEEKK HYLMWGLSSD FWGEKPNLSY
IIGKDTWVEH WPEEDECQDE ENQKQCQDLG AFTESMVVFG CPN

Signal peptide
C3 beta chain
C3 alpha chain/C3a anaphylatoxin
C3a alpha chain fragment
C3g fragment
C3d fragment
C3f fragment
C3c alpha chain fragment 2

Figure 4.10. Fragments of human complement component C3 precursor.
The sequence of human complement component C3 was retrieved from SwissProt. Specific fragments are represented in different colours and N-terminal Arg-C peptide sequences are underlined.

General information	
Entry name	CO3_HUMAN
Accession number	P01024
Integrated	21-JUL-1986, UniProtKB/Swiss-Prot.
Sequence update	12-DEC-2006, sequence version 2
Annotation update	10-JUN-2008, entry version 115
UniSave	P01024
UniRef100	UniRef100_P01024
UniParc	UPI000013EC9B

Key	Begin	End	Length	Description
SIGNAL	1	22	22	
CHAIN	23	1663	1641	Complement C3. /FTId=PRO_0000005907.
CHAIN	23	667	645	Complement C3 beta chain. /FTId=PRO_0000005908.
CHAIN	672	1663	992	Complement C3 alpha chain. /FTId=PRO_0000005909.
CHAIN	672	748	77	C3a anaphylatoxin. /FTId=PRO_0000005910.
CHAIN	749	1663	915	Complement C3b alpha' chain. /FTId=PRO_0000005911.
CHAIN	749	954	206	Complement C3c alpha' chain fragment 1. /FTId=PRO_0000005912.
CHAIN	955	1303	349	Complement C3dg fragment. /FTId=PRO_0000005913.
CHAIN	955	1001	47	Complement C3g fragment. /FTId=PRO_0000005914.
CHAIN	1002	1303	302	Complement C3d fragment. /FTId=PRO_0000005915.
PEPTIDE	1304	1320	17	Complement C3f fragment. /FTId=PRO_0000005916.
CHAIN	1321	1663	343	Complement C3c alpha' chain fragment 2. /FTId=PRO_0000273948.

Figure 4.11. SwissProt entry for Complement C3 precursor.
The entry was retrieved through SRS. The feature table shows how the component fragments *in vivo* to produce various polypeptides.

(Figure 4.11). The m/z values for the N-acetylated, Arg-C, N-terminal peptides were calculated using the Protein/Peptide Editor tool in MassLynx. The calculated m/z values were used to interrogate the Orbitrap data by preparing extracted ion chromatograms for the N-terminal of each complement fragment (Figure 4.12).

The most notable identification is that of the C3a anaphylatoxin N-terminal. As mentioned earlier, levels of this protein (which represents 5% of the parent C3 molecule; Hugli, 1975) have been shown to indicate the presence of colorectal tumours (Habermann *et al.*, 2006). Rapid and efficient identification of this protein in multiple human plasma samples could provide the basis of a diagnostic screen for this candidate biomarker. Furthermore, screening of high resolution MS data, using precise masses from other putative biomarkers, could provide the basis of a targeted approach to screen the preparation for other clinically interesting proteins.

4.5.6 Screen for unbound internal peptides

Due to the high abundance of proteins such as HSA in the sample, even trace amounts of internal peptides from these species could obscure the identification of lower abundance protein components in the mixture. It is possible to manually screen for internal peptides by preparing extracted ion chromatograms using m/z values from predicated internal Arg-C peptides. Figure 4.13 shows the predicted peptide sequences for an Arg-C digest of HSA. Extracted ion chromatograms were prepared using data from the Orbitrap mass spectrometer. When the intensity scale is fixed to that of the true N-terminal of HSA, the internal peptides are not found. Signals from peptides R4 and R21 were seen but at substantially lower levels than that of the true N-terminal peptide. In the case of R4 the ion is present at an intensity of 2.25E4 compared with 5.77E6 for the true N-terminal peptide ion. However, on closer inspection of the MS data it can be seen that the ion at m/z 661.327 is not clearly distinguishable as the $[M+3H]^{3+}$ ion and no MS/MS data was generated for this peptide (Figure 4.14). The peptide ion corresponding to R21 is present at an intensity of 7.72E5. In this case, the major species in the MS spectrum is the ion at m/z 697.847 corresponding to the R21 $[M+2H]^{2+}$ peptide ion. However, the product ion spectrum is not easily interpretable and for this reason it is not possible to confirm the sequence of this peptide (Figure 4.15).

This analysis shows that even in the case of HSA, which is by far the most abundant protein in human plasma, the method is not significantly 'leaky' it terms of the removal of internal peptides.

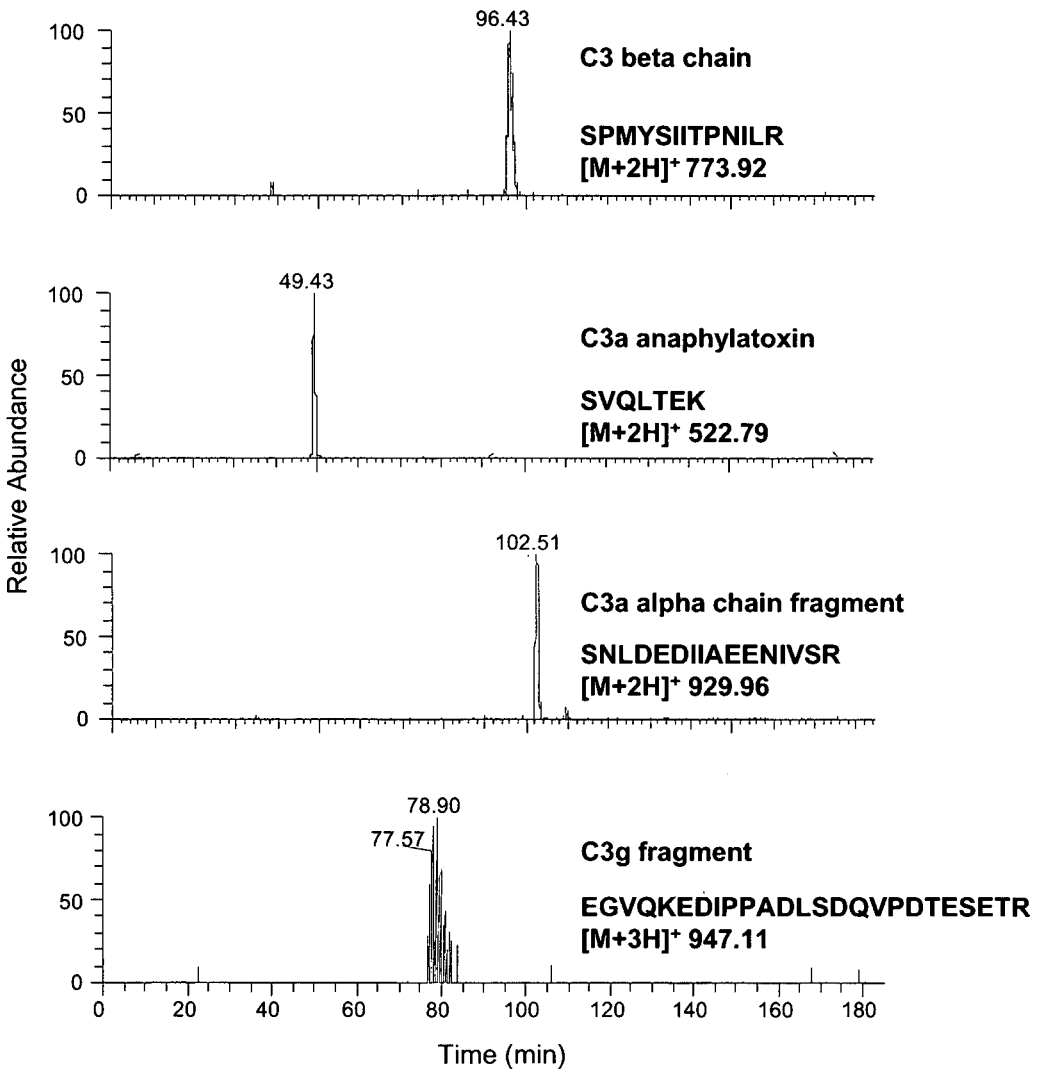


Figure 4.12. Manual analysis of complement component C3 fragment N-terminal sequences.
The N-terminal peptide preparation of human plasma was resolved by RP chromatography and analysed by high resolution ion trap MS on the Orbitrap instrument. The N-terminal Arg-C peptide was determined for each of the complement C3 derived fragments (m/z values determined using the Bioworks protein/peptide editor in MassLynx). Values within the mass range of the instrument were used to prepare extracted ion chromatograms.

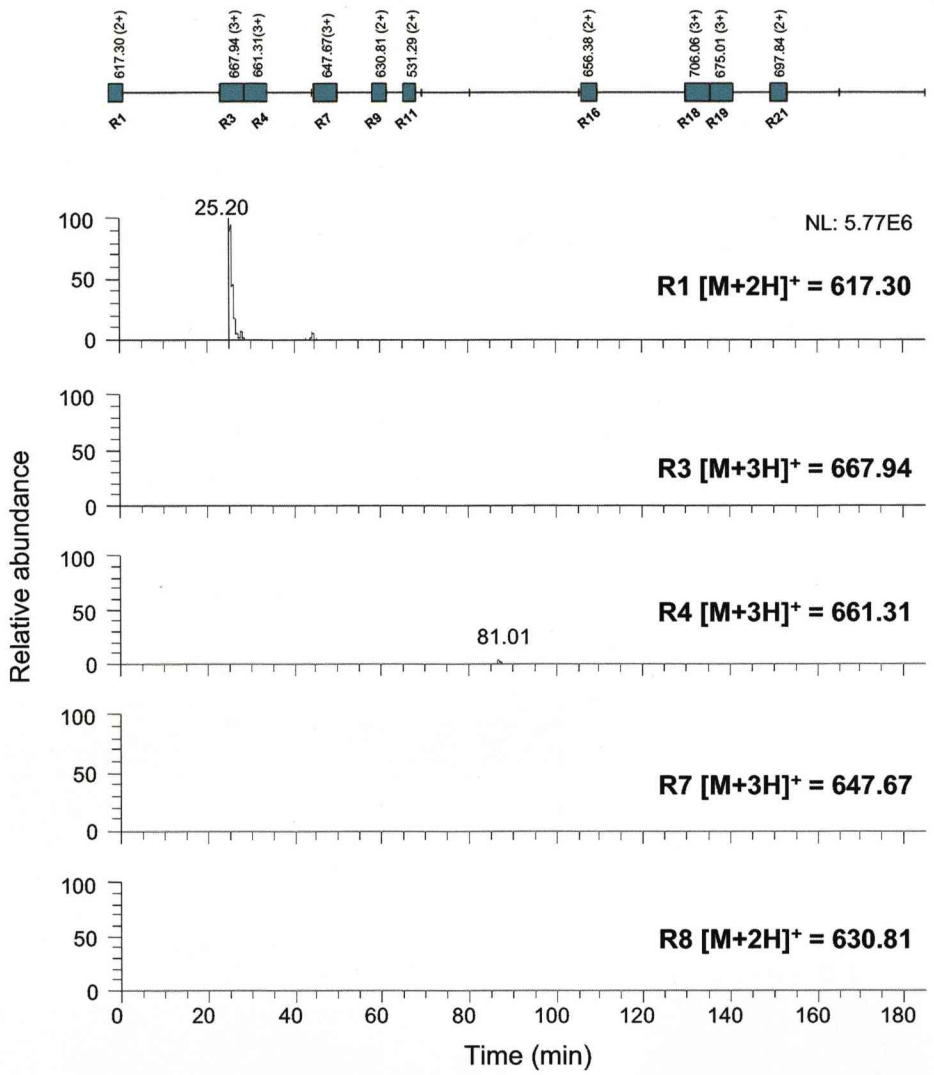


Figure 4.13. Screen for internal peptides from HSA.
An N-terminal peptide preparation of human plasma was resolved by RP chromatography and analysed by high resolution ion trap MS on the Orbitrap instrument. Extracted ion chromatograms were prepared for acetylated Arg-C peptides from human serum albumin (m/z values determined using the Bioworks protein/peptide editor in MassLynx). In all cases the intensity scale was fixed to that observed for the HSA mature N-terminal peptide (5.77×10^6).

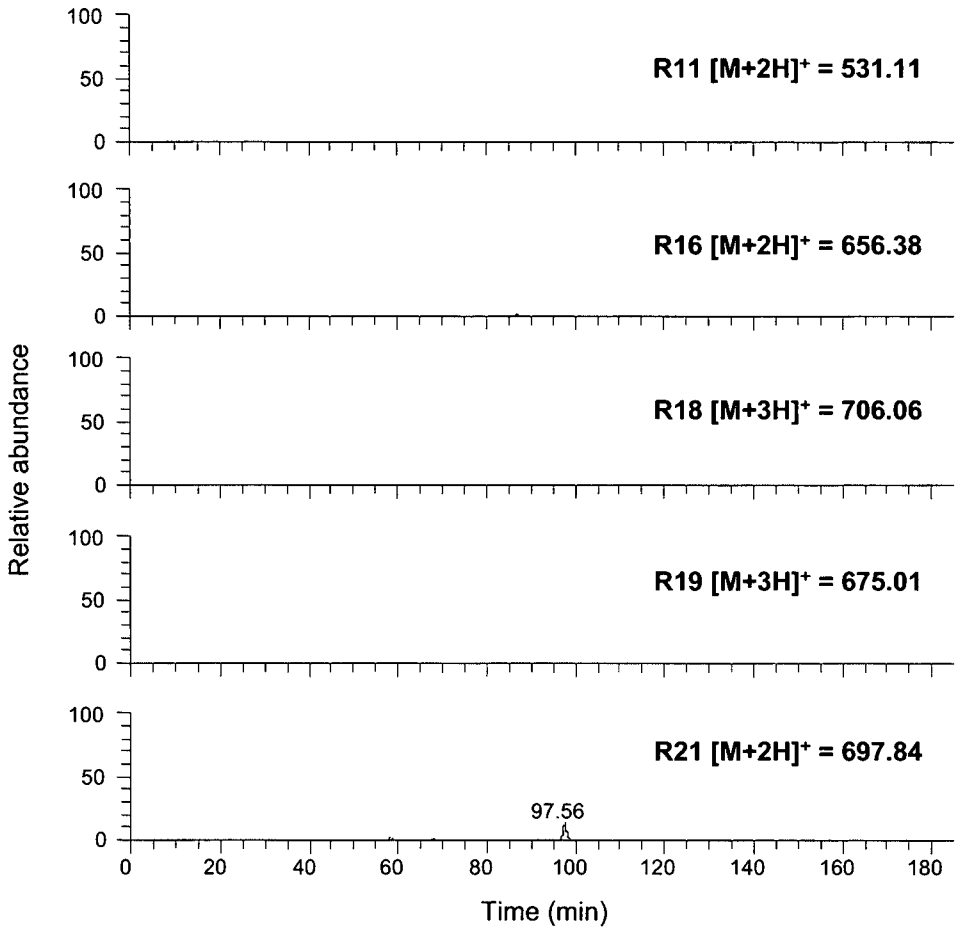


Figure 4.13. Screen for internal peptides from HSA. continued

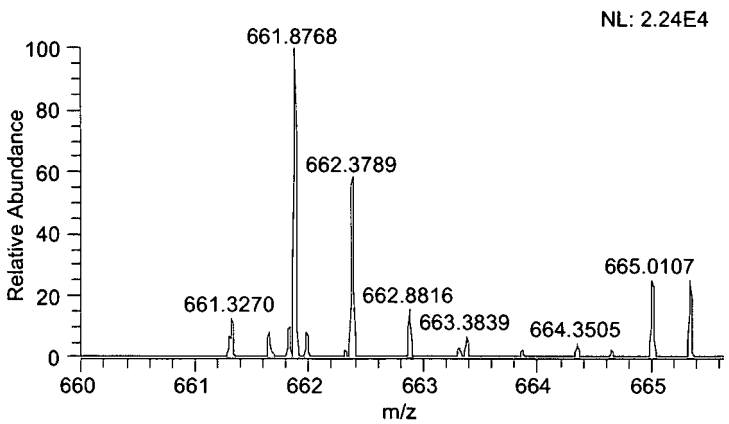


Figure 4.14. Presence of Arg-C peptide R4 from serum albumin in the human plasma N-terminal preparation.
The $[M+3H]^{3+}$ value of the internal Arg-C peptide R4 from HSA was used to prepare an extracted ion chromatogram on the LC-MS/MS data obtained from the Orbitrap analysis.

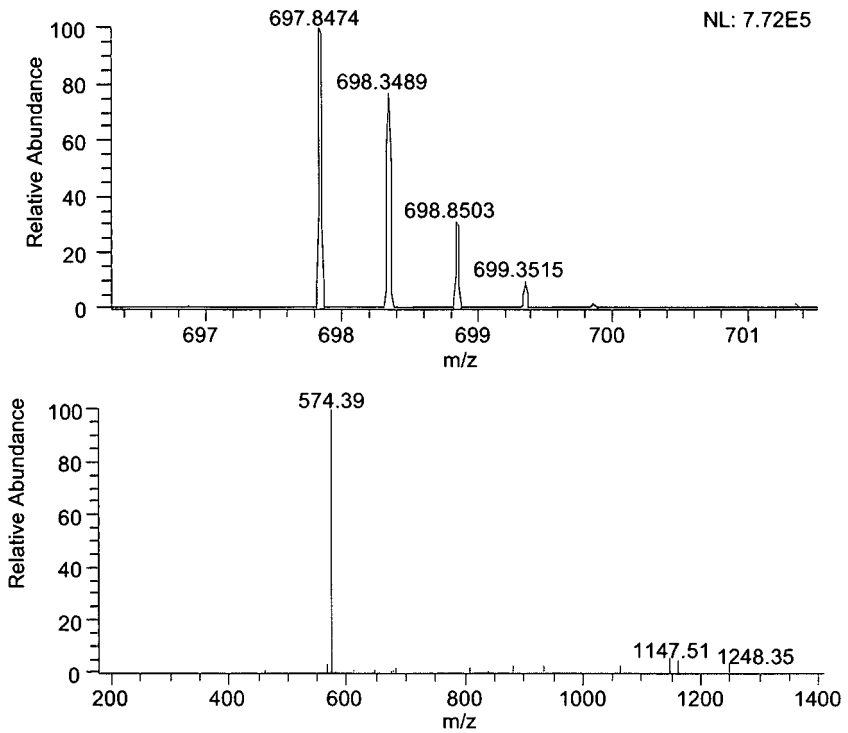


Figure 4.15. Presence of Arg-C peptide R4 from HSA in the plasma N-terminal preparation.
The $[M+2H]^{2+}$ value of the internal Arg-C peptide R21 from HSA was used to prepare an extracted ion chromatogram on the LC-MS/MS data obtained from the Orbitrap analysis (upper panel). The corresponding MS/MS fragmentation data was also extracted (lower panel).

4.5.7 Screen for truncated N-terminal peptides

N and C-terminal regions of proteins are subjected to a variety of exopeptidases that are capable of shortening the peptide chain by effectively cleaving or "trimming" individual or multiple amino acids in succession. Aminopeptidases catalyse the cleavage of amino acids from the amino terminus of protein or peptide substrates (Taylor, 1993) and carboxypeptidases catalyse the cleavage of amino acids from the C-terminus (Christianson and Lipscomb, 1989). These processes are common in plasma, as this fluid contains a large amount of protease (Redlitz *et al.*, 1995; Durinx *et al.*, 2000). Consequently, plasma proteins may exist in multiple forms consisting of varying degrees of truncation.

When searching the N-terminal MS/MS data using the standard Mascot parameters (listed in Section 2.11.3), the enzyme Arg-C is chosen, which limits the search space to arginine terminated, complete, N-terminal peptides. Alternatively, by altering the enzyme specificity to semi-trypsin, the search space is increased to allow for incomplete tryptic peptides (i.e. will allow for truncation at the N-terminus). In this case, N-terminal peptides that have been subjected to aminopeptidase cleavage will also be identified.

MS/MS data from the human plasma N-terminal preparation was searched against the human N-terminal data base, using the local Mascot server, with semi-trypsin as a parameter. The search identified multiple forms of N-terminal regions in many of the classical plasma proteins (data not shown). These multiple forms consisted of trimmed or shortened versions of the full length peptide from the N-terminal end of the protein. Low scores indicated that these shortened forms were present at lower intensities than the true N-terminal Arg-C peptides. Figure 4.16 shows extracted ion chromatograms for various forms of the C3a anaphylatoxin and serum amyloid N-terminal Arg-C peptide.

In all cases, the full length peptide is present at higher intensities than the N-terminally trimmed versions, assuming equal response factors. To assess the extent of this trimming process extracted ion chromatograms were prepared for various shortened forms of Arg-C N-terminal peptides from N-terminal preptides identified in Table 4.5. MS intensities for each peptide were recorded (Table 4.6) and used to plot the relative intensities (expressed as a percentage of the intensity of the true N-terminal Arg-C peptide) of the truncated forms (Figure 4.17).

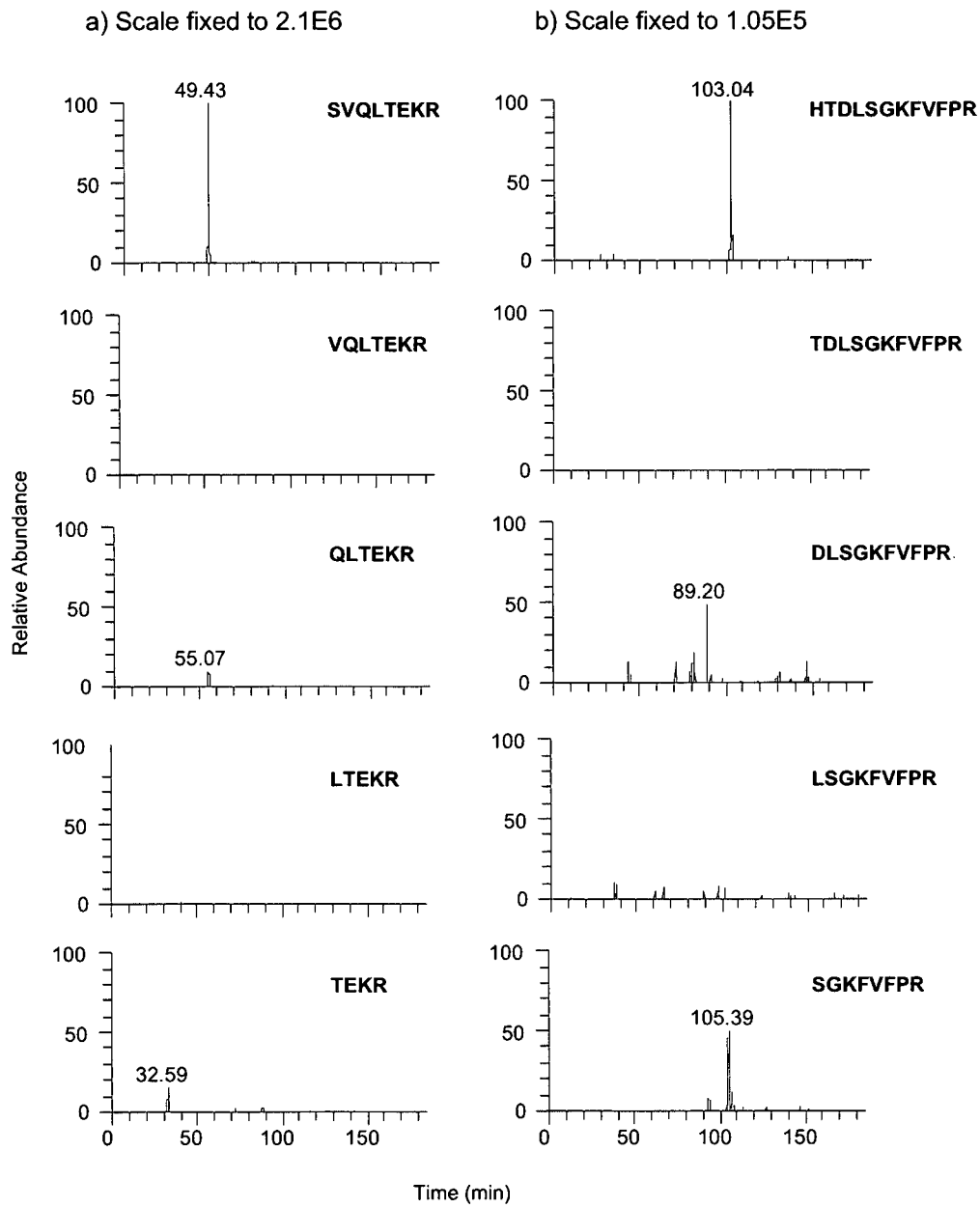


Figure 4.16. Identification of full length and truncated forms N-terminal peptides. The N-terminal peptide preparation of human plasma was resolved by RP chromatography and analysed by high resolution ion trap MS on the Orbitrap instrument. Extracted ion chromatograms were prepared for full length and truncated N-terminal Arg-C peptides from C3a anaphylatoxin and serum amyloid (m/z values determined using the Bioworks protein/peptide editor in MassLynx).

Protein	m/z	Charge	Intensity	Sequence
Serum albumin	638.30	2	$2.43x10^6$	DAHKSEVAHR
	580.79	2	$3.13x10^4$	AHKSEVAHR
	545.27	2	$5.02x10^5$	HKSEVAHR
	476.71	2	$7.85x10^4$	KSEVAHR
	740.37	2	0	SEVAHR
Serum amyloid	529.64	3	$1.09x10^5$	HTDLSGKFVFPR
	745.39	2	0	TDLSGKFVFPR
	695.87	2	$5.39x10^4$	DLSGKFVFPR
	637.35	2	0	LSGKFVFPR
	581.81	2	$5.05x10^4$	SGKVFPR
Fibrinogen alpha	789.85	2	$2.72x10^6$	ADSGEGDFLAEGGGVR
	754.34	2	$1.4x10^5$	DSGEGDFLAEGGGVR
	696.83	2	$1.03x10^5$	SGEGDFLAEGGGVR
	653.31	2	0	GEGDFLAEGGGVR
	624.8	2	$6.53x10^4$	EGDFLAEGGGVR
Alpha-1-anantichymotrypsin	1089.49	2	$2.9x10^5$	HPNSPLDEENLTQENQDR
	1020.96	0	0	PNSPLDEENLTQENQDR
	972.44	2	$7.22x10^4$	NSPLDEENLTQENQDR
	915.41	2	$6.46x10^3$	SPLDEENLTQENQDR
	871.90	2	0	PLDEENLTQENQDR
Carboxypeptiase B	664.86	2	$5.04x10^5$	FQSGQVLAALPR
	591.35	2	$1.70x10^4$	QSGQVLAALPR
	527.31	2	$4.34x10^4$	SGQVLAALPR
	483.79	2	0	GQVLAALPR
	455.28	2	0	QVLAALPR
Apolipoprotein C	901.45	3	$5.90x10^4$	TPDVSSALDKLKEFGNTLEDKAR
	867.77	3	0	PDVSSALDKLKEFGNTLEDKAR
	835.42	3	$1.83x10^4$	DVSSALDKLKEFGNTLEDKAR
	797.08	3	0	VSSALDKLKEFGNTLEDKAR
	764.06	3	0	SSALDKLKEFGNTLEDKAR
Alpha-2-HS-glycoprotein	561.81	2	$1.67x10^6$	APHGPGLIYR
	526.29	2	0	PHGPGLIYR
	477.76	2	$7.71x10^4$	HGPGLIYR
	817.46	1	0	GPGLIYR
	760.44	1	0	PGLIYR
C3a anaphylatoxin	522.79	2	$6.05x10^5$	SVQLTEKR
	957.54	1	0	VQLTEKR
	858.45	1	$8.02x10^4$	QLTEKR
	730.41	1	$6.76x10^3$	LTEKR
	617.31	1	$1.15x10^5$	TEKR

Table 4.6. Identification of N-terminally trimmed proteins by manual extraction of ions from the Orbitap acquisition of human plasma N-terminal peptides.
N-terminal Arg-C masses were determined for full and truncated forms of the major plasma proteins. Various m/z values were used to prepare extracted ion chromatograms from the Orbitrap LC-MS/MS analysis (three hour RP gradient) of the plasma N-terminal preparation.

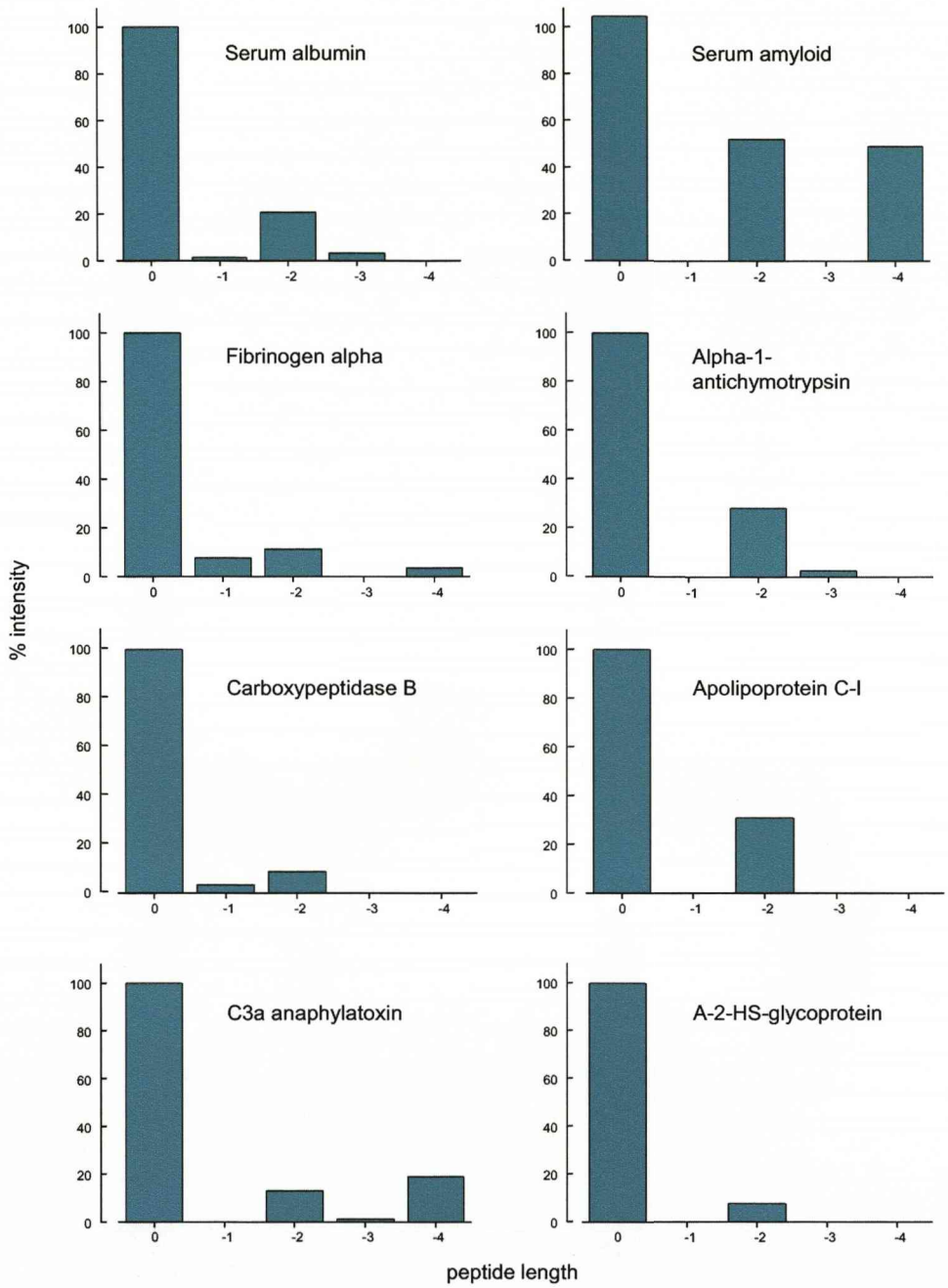


Figure 4.17. N-terminal trimming of human plasma proteins.
Extracted ion chromatograms were prepared using m/z values of complete sequences and trimmed forms of mature N-terminal peptides from the LC-MS/MS analysis on the Orbitrap instrument. The percentage intensities relative to the most intense species were plotted against the peptide length.

In most cases, the full length, mature, N-terminal peptide represents the most abundant species. However, in the case of serum amyloid, the C-2 peptide and the C-4 peptide are also present in high quantities (50% intensity of the unmodified N-terminal).

The existence of proteolytically trimmed N-termini in the preparation will undoubtedly increase the overall complexity of the sample. Even if these shortened forms are present in lower quantities, the presence of truncated versions of high abundant proteins in the sample will compromise signals from N-terminal peptides originating from low abundance species.

In five out of eight cases, there is evidence that trimming is caused by dipeptidyl peptidase activity, whereby two amino acids are cleaved together. Dipeptidyl peptidases are a group of enzymes capable of cleaving dipeptides from the N-terminal regions of proteins. Dipeptidyl aminopeptidase IV (DPP IV) was discovered in 1966 (Hopsu-Havu and Glenner, 1966) and is a serine peptidase which removes N-terminal dipeptides from polypeptides with penultimate proline and alanine residues (Kenny *et al.*, 1976). DPP IV is in close contact with proteins circulating in the blood, since it is located on endothelial cells (Lojda, 1979) and, more importantly, is found as a soluble enzyme in blood plasma (Durinx *et al.*, 2000). For more than 20 years, increased DPP IV activity has been linked to various diseases, including malignant processes (Hino *et al.*, 1976), rheumatoid arthritis (Hagihara *et al.*, 1987) and more recently diabetes mellitus (Ryskjaer *et al.*, 2006).

The classical plasma proteins: serum albumin, α-1-antichymotrypsin, apolipoprotein C and α-2-HS-glycoprotein each contain either alanine or proline at the penultimate position. Given the data presented in Figure 4.14, these proteins are all strong candidates for dipeptidyl peptidase activity. It may therefore be feasible to measure the activity of dipeptidyl peptidases by identifying and quantifying the substrates of these enzymes i.e. trimmed protein N-termini. The N-terminal positional proteomics strategy provides an ideal tool in which to measure the activity of these enzymes as it targets the processed region of the putative substrates.

4.5.8 Normalisation of plasma proteins

In an attempt to reduce the dynamic range of proteins in human plasma, Protein Equalizer™ technology was used to normalise the concentrations of individual components within the sample (a detailed description can be found in Section 1.5.5). For normalisation, 20mg of Equalizer™ beads were washed and swollen in 50% (v/v) methanol, prior to equilibration in 20mM Na_2CO_3 buffer. Human plasma proteins (100mg) were incubated with the beads for 2h,

after which beads were washed thoroughly to remove unbound protein. It was necessary to use a large excess of protein in order to saturate all of the potential binding sited on the ligand library beads. The washed bead suspension was subjected to 1-D SDS-PAGE along with 15µg of starting material (Figure 4.18). Wash fractions were also analysed to ensure that all unbound proteins had been removed (data not shown). From 1-D SD-PAGE analysis it can be seen that the dynamic range was successfully reduced in the plasma sample treated with the Equalizer™ beads. The two gel profiles are quite distinct, in the lane containing the starting material, the band at 66kDa corresponding to HSA dominates the gel and Coomassie staining fails to pick out many other protein bands. In contrast, the 1-D profile from the normalised sample contains many more bands visible by Coomassie staining and the gel lane is not dominated by one single band.

To establish if normalisation increases the amount of protein identifications obtained by in-solution digestion and LC-MS/MS analysis, the protein/bead suspension was subjected to tryptic digestion and the resulting peptides were analysed by RP LC-MS/MS, using an extended three hour gradient. Peptide sequences were searched against SwissProt using the Mascot search engine with taxonomy restricted to *Homo sapiens*. Proteins identified with a Mowse score of greater than 50 (ion score threshold) were accepted as confident (Table 4.7). Following normalisation 41 proteins were identified, which is a substantial improvement in comparison to the starting material, which generated only five significant hits (Table 4.4).

4.5.9 N-terminal tryptic peptide isolation of normalised plasma proteins

To investigate if normalisation in combination with the positional proteomics strategy would further improve the efficiency of protein identification, the N-terminal isolation protocol was performed on plasma proteins coupled to Protein Equalizer™ beads. The N-terminal protocol was performed directly on the beads without elution of proteins. The proteins were acetylated whilst coupled to the beads, subsequent proteolysis of the acetylated proteins liberated peptides from the beads. The resulting peptide mixture was removed from the beads in the supernatant fraction. N-terminal purification by NHS-activated Sepharose gave rise to a new set of N-terminal peptides. Figure 4.19 shows the N-terminal MALDI-ToF spectrum from the normalised human plasma preparation. The peak at 1233.6Da corresponding to the N-terminal of HSA, which dominates the N-terminal MALDI-ToF spectrum obtained from the starting material (Figure 4.6), is present but at a much lower intensity relative to the base peak ion (apolipoprotein A-I). Reduction of the intensity of the HSA N-terminal peptide, in turn,

Figure 4.18. Normalisation of protein concentrations in human plasma.
Human plasma (100mg, Sigma) was normalised using Protein Equalizer™ beads. The normalised protein/bead suspension (10μl) was separated by 1-D SDS-PAGE alongside 10μg of starting material.

	Protein	Accession	Mowse score	Coverage (%)	Peptides matched	Molecular weight
1	Apolipoprotein A-I	P02647	4461	61	28	30.76
2	Apolipoprotein A-IV	P06727	2796	32	16	45.37
3	Fibrinogen alpha chain	P02671	1647	13	12	94.91
4	Complement C4-B	P0C0L5	1631	5	6	192.64
5	Serum albumin	P02768	1113	19	13	69.32
6	Prothrombin	P00734	726	16	14	69.99
7	Vitronectin	P04004	380	9	5	54.271
8	Gelsolin	P06396	782	8	4	85.64
9	Inter-alpha-trypsin inhibitor heavy chain H1	P19827	191	3	2	101.32
10	Hemoglobin subunit beta	P68871	262	27	3	15.99
11	Apolipoprotein C-III	P02656	216	15	2	10.86
12	Ceruloplasmin	P00450	256	9	7	122.13
13	Clusterin	P10909	716	12	7	52.46
14	Apolipoprotein E	P02649	598	22	5	36.12
15	Apolipoprotein B-100	P04114	368	3	7	51.52
16	Ig lambda-1 chain C region	P15814	250	4	1	36.03
17	Fibronectin	P02751	209	7	13	262.44
18	Alpha-2-HS-glycoprotein	P02765	211	8	2	39.30
19	Complement C3	P01024	1341	6	11	187.03
20	Plasminogen	P00747	57	5	4	90.51
21	Apolipoprotein A-II	P02652	526	33	3	11.16
22	Complement C1q subcomponent subunit A	P02746	128	12	2	25.75
23	Inter-alpha-trypsin inhibitor heavy chain H1	Q14624	193	11	10	106.37
24	Fibrinogen beta chain	P02675	574	9	4	55.89
25	Fibrinogen gamma chain	P02679	289	13	3	51.48
26	Inter-alpha-trypsin inhibitor heavy chain H2	P19823	176	11	10	106.30
27	Complement factor H	P08603	173	3	3	123.98
28	Ig kappa chain C region	P01834	168	15	2	11.7
29	Hemoglobin subunit apha	P69905	51	10	1	15.24
30	Complement C1q subcomponent subunit B	P02746	117	11	2	26.42
31	Transthyretin	P02766	102	8	2	15.77
32	Plasminogen	P00747	100	6	4	90.10
33	Alpha-1-antitrypsin	P01011	75	25	3	46.70
34	Hemoglobin subunit delta	P02042	71	31	3	16.45
35	Complement C1q subcomponent subunit C	P02747	74	3	2	25.57
36	Complement C1s subcomponent	P09871	70	4	7	76.64
37	Pigment epithelium-derived factor	P36955	68	25	4	39.24
38	Complement component C8 alpha chain	P07357	64	5	2	67.00
39	Hemoglobin subunit epsilon	P02100	61	16	1	16.19
40	Complement component C8 beta chain	P07358	55	3	1	67.03
41	Apolipoprotein C-I	P02654	55	20	1	9.33

Table 4.7. Identification of normalised human plasma proteins by in-solution digestion and LC-MS/MS analysis.

The normalised bead/protein suspension (5µl) was subjected to proteolysis using trypsin (0.1µg). The supernatant containing the digested peptides was removed from the suspension and diluted 1 in 50 with 0.1% FA prior to LC-MS/MS analysis using a three hour RP gradient. MS/MS data was used to search the human N-terminal database using the MASCOT search engine. The taxonomy was restricted to *Homo sapiens;* fixed modifications: carbamidomethylation of cysteine; variable modification: oxidation of methionine; protease: trypsin; missed cleavages: 1; peptide tolerance: 1.5Da; MS/MS tolerance: 0.6Da; instrument: ESI-TRAP; peptide charge: 1+, 2+ and 3+. Protein identifications with a Mowse score greater than 50 were accepted as confident identifications.

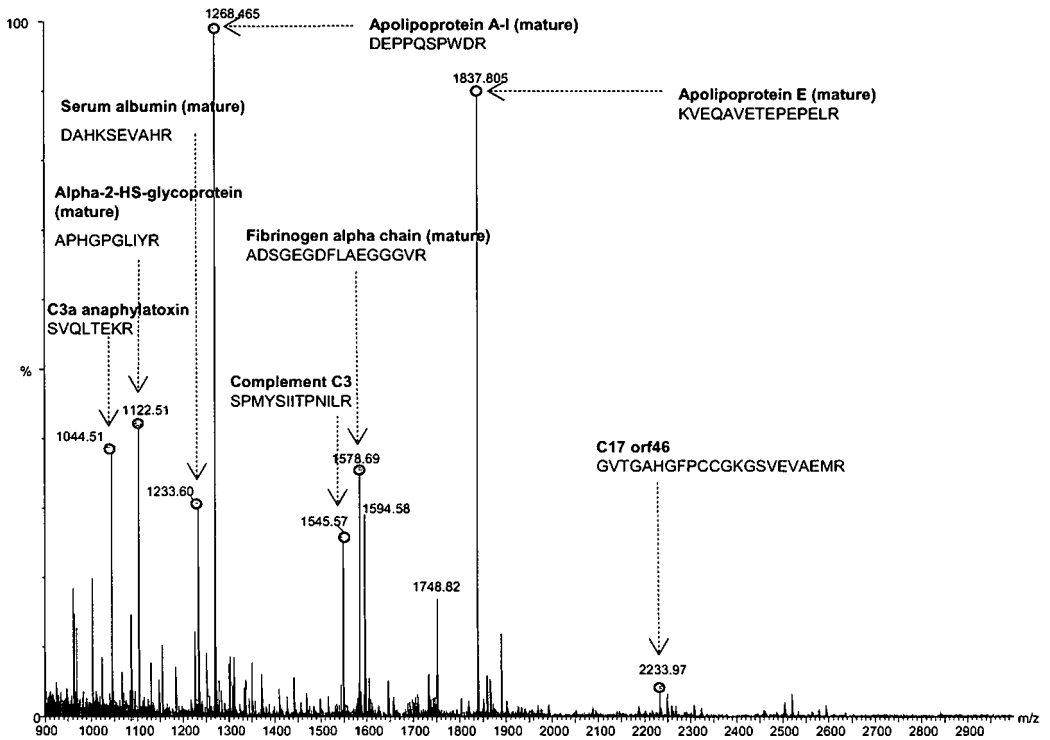


Figure 4.19. Normalised human plasma N-terminal preparation.
Normalised human plasma proteins (Sigma) coupled to Protein Equalizer™ beads were acetylated, washed and digested with trypsin. The peptides were separated from the bead suspension and incubated with 100µl NHS-activated Sepharose. A small portion (10µl) of the of supernatant was removed and desalted using a C18 ZipTip. The desalted peptides were analysed by MALDI-ToF MS. The major ions in the spectrum can be assigned to masses of acetylated, arginine terminated, true N-terminal peptides from plasma proteins.

allows signals from the other species in the spectrum to emerge. The cluster of peaks seen in the analysis from the starting material that arise from the multiple immunoglobulin isoforms is not visible in the normalised sample. New peaks at m/z 1044.5, 1545.6, 1837.8 and 2233.97 are present in the normailised MALDI-ToF N-terminal spectrum, corresponding to C3a anaphylatoxin, complement C3, apolipoprotein E and C17 orf46 respectively. Analysis of the normalised N-terminal preparation by LC-MS/MS using the ion trap instrument yielded a total of 55 protein identifications, 23 of which were not identified in the N-terminal preparation from untreated plasma (Table 4.8). A summary of proteins identified from human plasma, using in-solution digest and N-terminal purification of untreated plasma and normalised preparations can be seen in Table 4.9. A total of 102 unique protein identifications were made.

4.6 SUMMARY

This study demonstrates the use of a positional proteomics approach to characterise the true N-terminal region of proteins in human plasma. Although the use of N-terminal purification failed to deliver as comprehensive a survey of the plasma proteome compared with other methods (Omenn *et al.*, 2005; Shen *et. al.,* 2005), the approach has the added advantage of defining the true N-terminal peptide sequence from each protein. As highlighted in this study many plasma proteins undergo extensive processing at their N-terminal. This processing results in a variety of isoforms from each protein giving rise to multiple N-terminal sequences. Dipeptidyl peptidase activity appears to play a major role in N-terminal trimming. Many of the truncated forms identified were two amino acids shorter than the previously identified form. The implication of multiple N-terminal forms is increased complexity of the overall N-terminal preparation. Even though the truncated forms are present in lower quantities they will be significant enough to suppress the signal intensity of other less abundant proteins in the mixture. Identification and quantification of dipeptidyl peptidase substrates could provide a useful tool for measuring the activity of these enzymes which have roles in disease processes.

Application of the positional proteomics strategy to untreated human plasma resulted in the identification of over 20 N-terminal sequences corresponding to the variable region of immunoglobulin chains. Although this presents a problem in terms of global protein identification, the rapid identification of multiple immunoglobulin isoforms by N-terminal peptide isolation could have potential clinical implications in immunoglobulin profiling.

The use of Protein Equalizer™ technology to effectively normalise protein concentrations in the sample led to the identification of a new subset of proteins, in addition to

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score
1	lipopolysaccharide-binding protein	838.47	ANPGLVAR	SP	21
2	C6orf21 variant protein	865.36	ASSAMSDR	M	26
3	unknown	964.41	MKGMGSDR		41
4	Telomeric repeat-binding factor 2	1005.41	AGGGGSSDGSGR	M	28
5	DNA-binding protein inhibitor ID-2 (Q02363)	1019.53	MKAFSPVR		20
6	Caveolin-1	1024.43	SGGKYVDSE	M	29
7	Kruppel-like factor 6	1038.55	SAANPLTAPR	M	36
8	Protein C1orf97 (Q9BRT7)	1127.54	MDKKSTHR		39
9	Mitochondrial ribosomal protein S24	1171.60	AASVCSGLLGPR	M	48
10	Puromycin-sensitive aminopeptidase	1298.68	MWLAAAAPSLAR		51
11	Nuclear receptor coactivator 2	1320.53	SGMGENTSDPSR	M	20
12	Cyclic AMP-dependent transcription factor ATF-5 (Q9Y2D1)	1347.75	SLLATLGLELDR	M	31
13	GranzymeK (P49863)	1449.76	EIIGGKEVSPHSR	SP	34
14	Complement C5 (P01024)	1545.82	SPMYSIITPNILR	SP	45
15	meiotic nuclear divisions 1 homolog	1570.84	SKKKGLSAEEKR	M	25
16	Cholinesterase 9	1696.9	EDDIIIATKNGKVR	SP	22
17	potassium large conductance calcium-activated channel	1717.74	ANGGGGGGGSSGGGGGGGGSSLR	M	30
18	GRAM domain-containing protein 1B (Q3KR37)	1726.84	MKGFKLSCTASNSNR		36
19	EPC1 protein	1764.77	MEKEEESEHHLQR		45
20	B-cell linker protein (Cytoplasmic adapter protein)	2009.10	MDKLNKITVPASQKLR		32
21	Uncharacterized protein C17orf46 (Q96LK8)	2233.01	GVTGAHGFPCCGKGSVEVAEMR	M	40
22	Beta-sarcoglycan	2269.11	AAAAAAAAEQQSSNGPVKKSMR	M	20
23	alpha-1-antitrypsin	2774.14	EDPQGDAAQKTDTSHHDQDHPTFNR	SP	55

Table 4.8. N-terminal identifications from normalised human plasma proteins.
The N-terminal peptide preparation of normalised human plasma was analysed by LC-MS/MS using a three hour reverse-phase gradient. MS/MS data was used to search the human N-terminal database using the MASCOT search engine. The taxonomy was restricted to *Homo sapiens;* fixed modifications: N-terminal acetylation and lysine acetylation; variable modification: oxidation of methionine; protease: Arg-C; missed cleavages: 1; peptide tolerance: 1.5Da; MS/MS tolerance: 0.6Da; instrument: ESI-TRAP; peptide charge: 1+, 2+ and 3+. Protein identifications with a Mowse score greater than 20 were accepted as confident identifications.

	Protein	Untreated		Normalised	
		Total	NT	Total	NT
1	Apolipoprotein A-I	*	*	*	*
2	Apolipoprotein A-IV		*	*	*
3	Fibrinogen alpha chain		*	*	*
4	Complement C4-B			*	
5	Serum albumin	*	*	*	*
6	Prothrombin			*	
7	Vitronectin			*	
8	Transferrin	*	*	*	
9	Inter-alpha-trypsin inhibitor heavy chain H1			*	
10	Hemoglobin subunit beta		*	*	*
11	Apolipoprotein C-III			*	
12	Ceruloplasmin			*	
13	Clusterin			*	
14	Apolipoprotein E			*	
15	Apolipoprotein B-100			*	
16	Ig lambda-1 chain C region			*	
17	Fibronectin			*	
18	Alpha-2-HS-glycoprotein		*	*	*
19	Complement C3		*	*	*
20	Plasminogen			*	
21	Apolipoprotein A-II			*	
22	Complement C1q subcomponent subunit A			*	
23	Inter-alpha-trypsin inhibitor heavy chain H1			*	
24	Fibrinogen beta chain			*	
25	Fibrinogen gamma chain			*	
26	Inter-alpha-trypsin inhibitor heavy chain H2			*	
27	Complement factor H			*	
28	Ig kappa chain C region			*	
29	Hemoglobin subunit apha		*	*	*
30	Complement C1q subcomponent subunit B			*	
31	Transthyretin			*	
32	Plasminogen			*	
33	Alpha-1-antitrypsin		*	*	*
34	Hemoglobin subunit delta			*	
35	Complement C1q subcomponent subunit C			*	
36	Complement C1s subcomponent			*	
37	Pigment epithelium-derived factor			*	
38	Complement component C8 alpha chain			*	
39	Hemoglobin subunit epsilon			*	
40	Complement component C8 beta chain			*	
41	Apolipoprotein C-I		*	*	*
42	Transmembrane protein 163				
43	C3a anaphylatoxin		*		*
44	Mitochondrial import receptor				
45	antigen MLAA-39				
46	Glutathione S-transferase Mu 1				*
47	Oxidoreductase HTATIP2				
48	Antithrombin-III				
49	antigen NY-CO-43				
50	Transcription factor ETV6				

Table 4.9. Total protein identifications from human plasma.
Identifications from the in-solution tryptic digest of the starting material (untreated plasma) and normalised samples and identifications from the N-terminal preparations of the starting material and normalised samples were collated.

	Protein	Untreated		Normalised	
		Total	NT	Total	NT
51	Inter-alpha-trypsin inhibitor		*		
52	FBXW12 mature		*		
53	Tyrosin kinase receptor erbB-3		*		*
54	cyclin G associated kinase		*		*
55	Alpha-1-antichymotrypsin		*		*
56	Guanylate kinase		*		*
57	Replication protein A 30 kDa subunit		*		*
58	HIV-1 like protein		*		*
59	lipopolysaccharide-binding protein				*
60	C6orf21 variant protein				*
61	unknown				*
62	Telomeric repeat-binding factor 2				*
63	DNA-binding protein inhibitor ID-2				*
64	Caveolin-1				*
65	Kruppel-like factor 6				*
66	Protein C1orf97				*
67	Mitochondrial ribosomal protein S24				*
68	Puromycin-sensitive aminopeptidase				*
69	Nuclear receptor coactivator 2				*
70	lipopolysaccharide-binding protein				*
71	Cyclic AMP-dependent transcription factor ATF-5				*
72	GranzymeK				*
73	Complement C5				*
74	meiotic nuclear divisions 1 homolog				*
75	Cholinesterase 9				*
76	potassium large conductance calcium-activated channel				*
77	GRAM domain-containing protein 1B				*
78	EPC1 protein				*
79	B-cell linker protein (Cytoplasmic adapter protein)				*
80	Uncharacterized protein C17orf46				*
81	Beta-sarcoglycan				*
82	immunoglobulin heavy chain variable region		*		
83	immunoglobulin heavy chain variable region		*		
84	Ig heavy chain V-III region CAM		*		
85	Ig heavy chain V-III region GA		*		
86	Ig heavy chain V-III region BRO		*		
87	Ig kappa chain V-I region BAN		*		
88	Ig kappa chain V-I region Mev		*		
89	Ig kappa chain V-I region Wes		*		
90	Ig heavy chain VHDJ region		*		
91	Ig lambda chain V-III region SH		*		
92	Ig kappa chain V-I region DEE		*		
93	Ig heavy chain V-III region LAY		*		
94	Ig heavy chain V-III region BRO		*		
95	Ig kappa chain V-III region SIE		*		
96	Ig kappa chain V-I region CAR		*		
97	Ig heavy chain V-III region TIL		*		
98	Ig kappa chain V-III region VG mature		*		
99	Ig heavy chain V-III region JON		*		
100	Ig heavy chain V-III region BUT		*		
101	Ig kappa chain V-IV region Len		*		
102	Ig heavy chain V-III region TRO		*		

the proteins already identified from the N-terminal analysis of untreated plasma. Identification of the colorectal tumour biomarker C3a anaphylatoxin was enhanced in the ligand library treated plasma sample (Figure 4.19).

5: MASS ISOTOPE DISTRIBUTION ANALYSIS OF AMINO ACID RESIDUES

5.1 INTRODUCTION

5.1.1 Stable isotopes

Stable isotope labelling is an important tool in proteomics and is becoming increasingly popular (reviewed in Beynon and Pratt, 2005). In general, the introduction of a stable isotope does not affect the physicochemical properties of a protein, with differentially labelled species behaving the same during processes such as cell growth, proteolysis, chromatographic separation and fragmentation. Only at the final stage of analysis (MS), can they can be distinguished. Stable isotope labelling *in vivo* has mainly focused on comparative proteomic methods, in which the incorporation of a metabolic label is used to identify one of the components in a pair wise comparison (Blagoev *et al.*, 2003; Everley *et al.*, 2004; Ong *et al.*, 2002). Methods for the incorporation of stable isotopes in *vitro* include ICAT (Gygi, *et al.*, 1999; Shiio and Aebersold, 2006), [^{18}O] labelling during proteolysis (Reynolds *et al.*, 2002; Yao *et al.*, 2001) and guanidination (Beardsley and Reilly, 2002; Brancia *et al.*, 2001; Thevis *et al.*, 2003).

Incorporation of a labelled amino acid can also be used to provide additional information for both *de novo* sequencing and database searching, for instance, in counting the number of amino acids in a given peptide. If the frequency of a specific amino acid is readily apparent from manual inspection of the MS data, it is possible to incorporate this information as an additional parameter in the database search. The MASCOT search engine has the capability to allow the inclusion of amino acid composition data using appropriate search terms. For example, the string "[M+H] + comp (X[n])" not only submits the peptide mass to the search, but permits the addition of limited compositional information. Knowledge of the frequency (n) of a specific amino acid (x), significantly restricts the amount of search space required (Engen *et al.*, 2002; Hunter *et al.*, 2001; Martinovic *et al.*, 2002; Pan *et al.*, 2003; Zhu *et al.*, 2002; Pratt *et al.*, 2002). A separate approach utilising metabolic labelling (both [^{15}N] and [^{13}C]) has been adopted to determine the carbon and nitrogen composition of peptides (Snijders *et al.*, 2005).

5.1.2 Mass isotope distribution analysis

Mass isotope distribution analysis (MIDA) is a technique traditionally used for measuring the synthesis of biological polymers. The technique involves the quantification (by MS) of the relative abundance of molecular species of a polymer, differing in mass only (mass isotopomers), following the introduction of a stable isotope. Mass isotopomers differ by the number of heavy atoms in their molecules, e.g., [$^{14}N_2$] and [$^{15}N_2$] urea; or [H_2] and [$^2H_3$] acetic anhydride.

The polymers measured can range from simple fatty acids synthesised from acetyl-CoA units to complex protein molecules. MIDA analysis is based on the analysis of numerical distributions, therefore it is essential that different combinatorial possibilities exist for the molecule analysed (Hellerstein and Neese, 1992).

5.1.3 Development of a novel acetylation reagent

When developing the N-terminal positional proteomics strategy (Chapter 1), acetylation using acetic anhydride or sulfo-NHS acetate was chosen as the amino group blocking reagent because it converges, in terms of downstream chemistry, naturally acetylated peptides with those acetylated chemically as part of the protocol. However, a stable isotope labelled variant of acetic anhydride $(C[^2H_3]CO)_2O$ was also used to discriminate naturally N^{α}-acetylated peptides from those that are acetylated chemically. During the use of this reagent a minor ion was observed, 1Da smaller than the monoisotopic ion, corresponding to the chemically acetylated peptide (Figure 5.1). This feature is due to a trace (circa 1%) of $(C[^2H_2]CO)_2O$ in the reagent preparation, which also labels α and ε amino groups in peptides, but with a 1Da mass difference. The "minus 1" ion was not evident in naturally acetylated peptides that lacked lysine residues. It is possible to exploit this property of the reagent in order to identify the occurrence of a chemical acetylation site on each peptide in a mixture, which in turn provides information for identification of proteins by virtue of the N-terminal peptide. This chapter describes the design of a new reagent, comprising $(C[^2H_3]CO)_2O$ and $([^{13}C]H_3[^{13}C]O)_2O$ mixed in a 9:1 ratio. The relative ion intensity of the "minus 1" ion reflects the number of acetylated sites in the protein, which allows the number of amino groups (lysine residues) to be counted. This reagent, referred to as MIDAR (mass isotope distribution analysis of amino acid residues) introduces the principle of using mixed isotope tags to gain further information for protein identification and in particular, to count the frequency of specific amino acids.

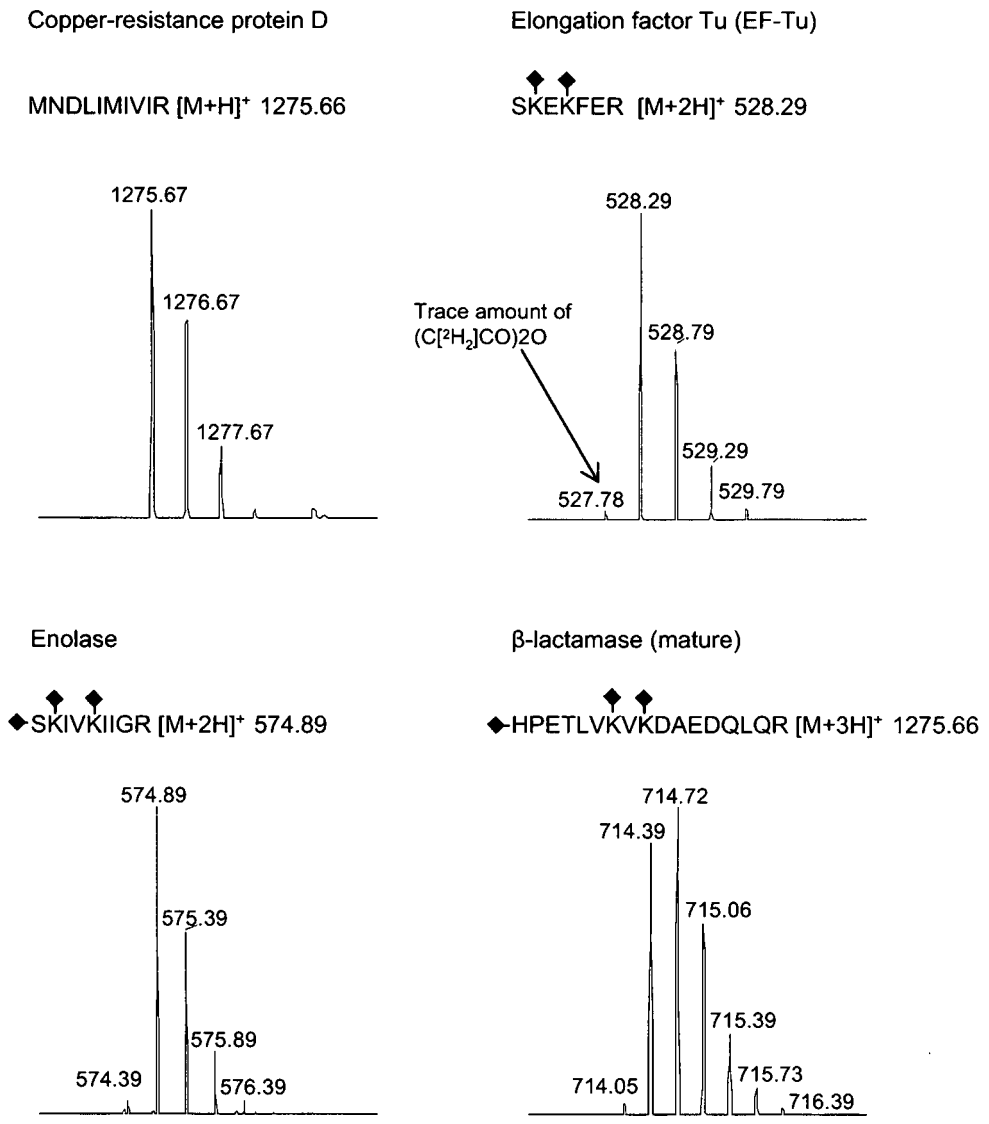


Figure 5.1. Observation of a slight impurity peak in $(C[^2H_3]CO)_2O$ labelled peptides.
The $(C[^2H_3]CO)_2O$ labelled, *E. coli* N-terminal preparation was analysed on the Orbitrap mass spectrometer by RP LC-MS/MS using an extended, three hour, gradient. Extracted ion chromatograms were prepared for N-terminal peptides containing 0, 2 and 3 acetyl groups.

The logic behind the design and use of the reagent is straightforward (Figure 5.2). Consider a peptide with n free amino groups (α and ε). If the acetic anhydride (of mass offset O) is a pure reagent, then complete reaction generates a single mass shift such that the product will have the mass increment of n.O Da. However, if the reagent is an isotopically coded mixture where the proportion of the heavier reagent is P (for example, P=0.9, 90% O=3, 10% O=2, the properties of the reagent used here) the resulting spectra are more complex. For a peptide with a single amino group (n=1) there are two products M+2Da and M+3Da in a 10:90 ratio, where M is calculated as the mass of the peptide after modification with acetic anhydride showing the natural isotope distribution (+42Da).

When a peptide contains more than a single amino group, the products become more complex. For instance, a peptide with three amino groups will, when reacted, generate four products M+(3*3)=M+9Da, (M+2*3+1*2)=M+8Da, (M+1*3+2*2)=M+7Da, (M+3*2)=M+6Da. Two of the species are labelled uniformly with either the O=2 reagent M+6Da or the O=3 M+9Da reagent, but the middle two (O=2,O=2,O=3, M+7Da) and (O=2,O=3,O=3, M+8Da) are characterised by mixed labelling patterns. Additionally, because of the positional combinations, there are three different ways to obtain peptides of these mass offsets (M+7Da: [+2,+2,+3], [+2,+3,+2] and [+3,+2,+2] for the lighter intermediate, and (M+8Da: [+2,+3,+3], [+3,+3,+2] and [+3,+2,+3] for the heavier) and their intensities are higher, as predicted by a binomial expansion (1:3:3:1 in this instance, where there are three labelling sites). The intensity ratio of the four ions is therefore dictated by the relative proportion of the two variants of the reagent. At P=0.9, the intensities will be in the proportions [M+9]: 72.9%, [M+8]: 24.3%, [M+7]: 2.7%, [M+6]: 0.1%. The combined spectrum, encompassing the reagent variants and the natural isotope abundance of each of the variants, will be complex to calculate, but at 10% O=2, the distinction between n=1, 2, 3 is clear without detailed analysis (Figure 5.3).

Further complexity in the mass spectrum for each labelled peptide derives from the combinatorial nature of the labelling by each form of the reagent (O=2, O=3). For a peptide with a single amino group (whether the α amino group, or an internal ε amino group derived from a lysine residue in a peptide that is blocked by natural acetylation), there can only be two variants, modified as O=2, and as O=3. Thus, the M+2Da ion will have an intensity in the mass spectrum that is approximately 10% of the M+3Da ion. The relative ion intensity is only approximately 10% because the mass spectrum is now the composite of two overlapping envelopes (monoisotopic mass, first [^{13}C] ion, second [^{13}C] ion etc for each of the M+2Da and M+3Da modified peptides). As the mass of the peptide increases, and as the number of amino

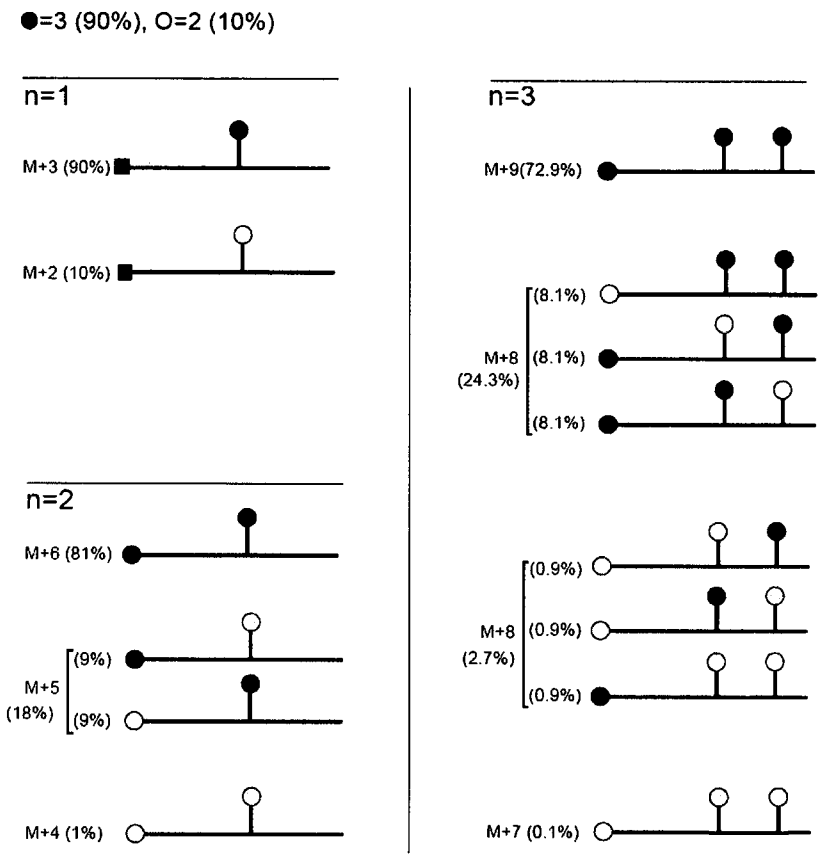


Figure 5.2. General principle of MIDAR.
Peptides containing different numbers of amino groups will be modified by the two isotope-coded (black circles, mass offset ●=3, light circles, mass offset **O**=2) variants of acetic anhydride. For peptides containing more than one amino group (**n**=2, **n**=3 in the figure), the two isotopic variants will be incorporated according to the combinatorial probabilities of the two variants. The intensities of the different mass variants is dictated by the relative proportion **P** of the two variants; in this example, the calculated intensities (in parentheses, as a percentage of the total intensity) are calculated for **P**(●=3)=0.9 and **P**(**O**=2)=0.1. In this illustration, the peptide illustrated for **n**=1 is provided by an N-terminal peptide that is naturally acetylated (closed squares), and which is therefore unavailable for further modification.

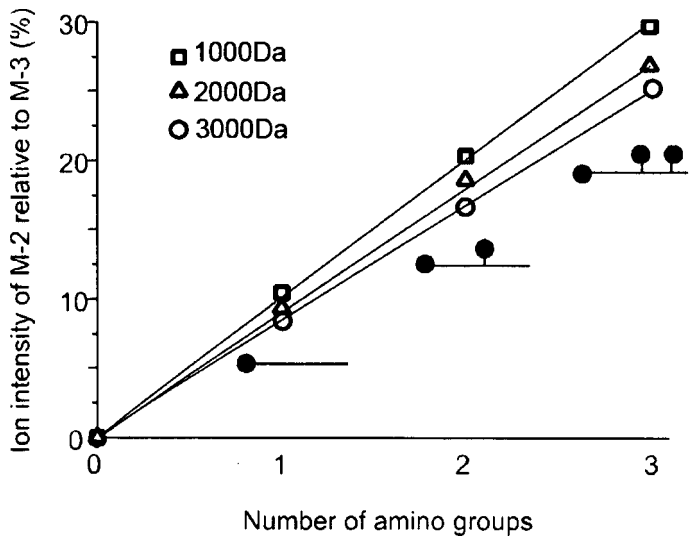


Figure 5.3. The effect of number of amino groups and peptide mass on the MIDAR isotope profile.
A series of theoretical peptides were created using an 'averagine' composition, at 1000Da, 2000Da and 3000Da, and the natural isotope distributes were calculated using the MS-Isotope program (described in Chapter 2). This distribution was then used to simulate the total isotopic profile, based on the combinatorial probabilities for incorporation of two variants of isotope-coded acetic anhydride. For each peptide mass, and for different numbers of amino groups, these isotope profiles were then used to calculate the intensity of the ion 1Da less than the monoisotopic mass (the 'minus-1 value').

groups also increases, the deviation becomes larger, but at no time does it compromise the simple assessment of the number of modified sites by visual inspection of the intensity of the "minus 1" ion and the monoisotopic ion of the species labelled with O=3. The incorporation of the doubly-labelled acetic anhydride is readily apparent from the highly atypical presence of the "minus 1" ion in the mass spectrum of each peptide.

5.1.4 Accurate mass and retention time

For simple peptide mixtures, there is increasing interest in the use of accurate mass tags, coupled with precise elution times, as sufficient parameters for proteome profiling (AMT strategy; Conrads *et al.,* 2000; Lipton *et al.,* 2006; Liu *et al.,* 2007). HPLC, which is used primarily as a simplification device, can be exploited to provide additional information regarding the properties of the species eluted from the chromatographic media. Although the resolving capabilities of HPLC and modern MS are not comparable, such information can be used to differentiate between peptides present in a complex mixture and more importantly, to provide additional information for protein/peptide identification. Models that are used to predict retention time are based on the assumption that the chromatographic behaviour of a peptide is dependent on its amino acid composition. However, modification of a peptide will affect the rate at which it is eluted from the reverse phase media by altering the physiochemical properties of the amino acid residues affected.

Protein modification by acetylation is heavily used in the strategies described in this thesis. In order to employ an AMT based approach to N-terminal identification, it is necessary to investigate the effect of acetylation on peptide retention time.

5.2 Aims and objectives

This aim of this chapter is to demonstrate the application of the MIDAR reagent to report the number of lysine residues contained within a peptide. The reagent will be used to modify various model peptides, purified proteins and naturally occurring proteins in order to establish its reproducibility. The reagent will then be used to replace unlabelled acetic anhydride as the acetylation reagent in the N-terminal positional proteomics strategy. This reagent is of particular interest in studying proteins from prokaryotes as these proteins are, in general, unmodified at the N-terminus. When used in higher eukaryotes, the majority of proteins will be

blocked by N^{α}-acetylation, therefore, modification by MIDAR is restricted to lysine residues. For these reasons, the soluble protein fraction of *E. coli* cell lysate will be the sample of choice to demonstrate MIDAR labelling of N-terminal peptides. Cells (*E. coli* K12) grown on both normal medium and an isotope-depleted medium ([^{13}C] and [^{15}N] deficient) were used in the analysis. The rationale behind the isotope depleted medium is simplification of the complex overlapping sets of naturally occurring isotope distributions seen in composite spectra of the differentially labelled variants.

Acetylation alters the physicochemical properties of proteins and peptides and will subsequently affect the relative hydrophobicity of a given peptide sequence. This modification will make retention time prediction using standard methods more challenging. An additional aim to this chapter is to investigate the extent to which acetylation affects the retention time of a set of peptides generated from the acetylation of a standard tryptic digest and an N-terminal preparation of a complex proteome. Once determined, this alteration may be used to suggest ways in which to predict the standard retention time of acetylated NTPeps.

5.3 Results and discussion

5.3.1 Labelling pattern

To assess labelling patterns, the model peptide ACTH18-39, sequence RPVKVYPNGAEDESAEAFPLEF, m/z 2465.13 was modified using the MIDAR reagent. The peptide, which contains two potential acetylation sites (the free N-terminal and side chain of the lysine residue), was reacted incompletely with the reagent to generate a mixture of single and doubly acetylated products (Figure 5.4). Two products were generated, at the expected values of m/z 2510.30 and m/z 2555.34 based on an increase in mass of 45Da and 90Da. At the same time, additional ions, not apparent in the unmodified peptide, were observed at 2509.32 m/z and 2554.32 m/z, consistent with the incorporation of both [$^{13}C_2$] and [$^{2}H_3$] acetyl groups. The intensities of these two additional "minus 1" ions were, relative to the monoisotopic peak, 11.5% and 22.4%, compared to the predicted values of between 9% and 10% for a single amino group, and between 17% and 19% for two amino groups (Figure 5.3). The slightly higher values might reflect the fact that the reagent was slightly more than 10% [$^{13}C_2$] acetic anhydride. Even so, it is readily apparent that even on visual inspection of the mass spectra of the modified peptides, it should be possible to assign the number of modified amino groups to each species.

5.3.2 Acetylation of model peptides and proteins

To determine the variance in the intensity of the isotope depleted ion, the same analysis was performed on a range of model peptides (Figure 5.5), purified proteins (Figure 5.6) and proteins from a real biological sample (mouse skeletal muscle; Figure 5.7 and 5.8). For all samples, the acetylated products were analysed by MALDI-ToF MS to examine the isotope distribution. In all instances, the number of acetylation sites was known, and ranged from one to five (Sequence details of model peptides and proteins can be found in Chapter 2). For each peptide, the intensity (peak height) of the "minus 1" ion was calculated as a percentage of the monoisotopic ion intensity (Figure 5.9). The correlation between the number of amino groups and the intensity ratio was 0.988 ($P<0.0001$). The slope of the line (approx. 11.5) suggested that the proportion of the lighter isotope variant in the preparation was slightly greater than 10%. This has no impact on the efficacy of the method, as the reagent is readily 'titrated' with model peptides. The slightly higher value is assumed to reflect the small percentage of the lighter species that was already present in the heavier variant of the reagent – the same

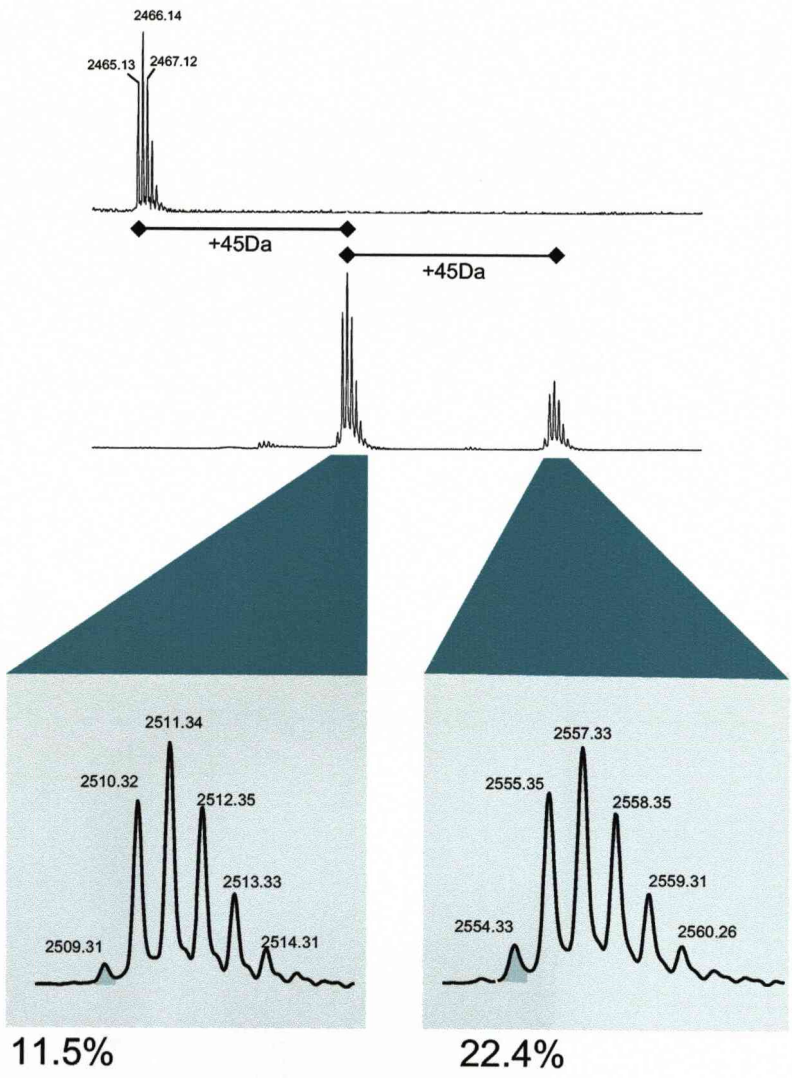


Figure 5.4. MIDAR profile for ACTH 18-39.
ACTH fragment 18-39 (RPVKVYPNGAEDESAEAFPLEF) was deliberately partially acetylated using combinatorially-coded acetic anhydride (**P**(**O**=3)=0.9 and **P**(**O**=2)=0.1) and the modified peptides were analysed by MALDI-ToF MS. The mass spectra of the two products with mass offsets of 45Da are apparent, corresponding to the singly-acetylated and doubly-acetylated forms of the peptide. For each, the intensity of the "minus 1" ion (shaded) was calculated relative to the monoisotopic intensity.

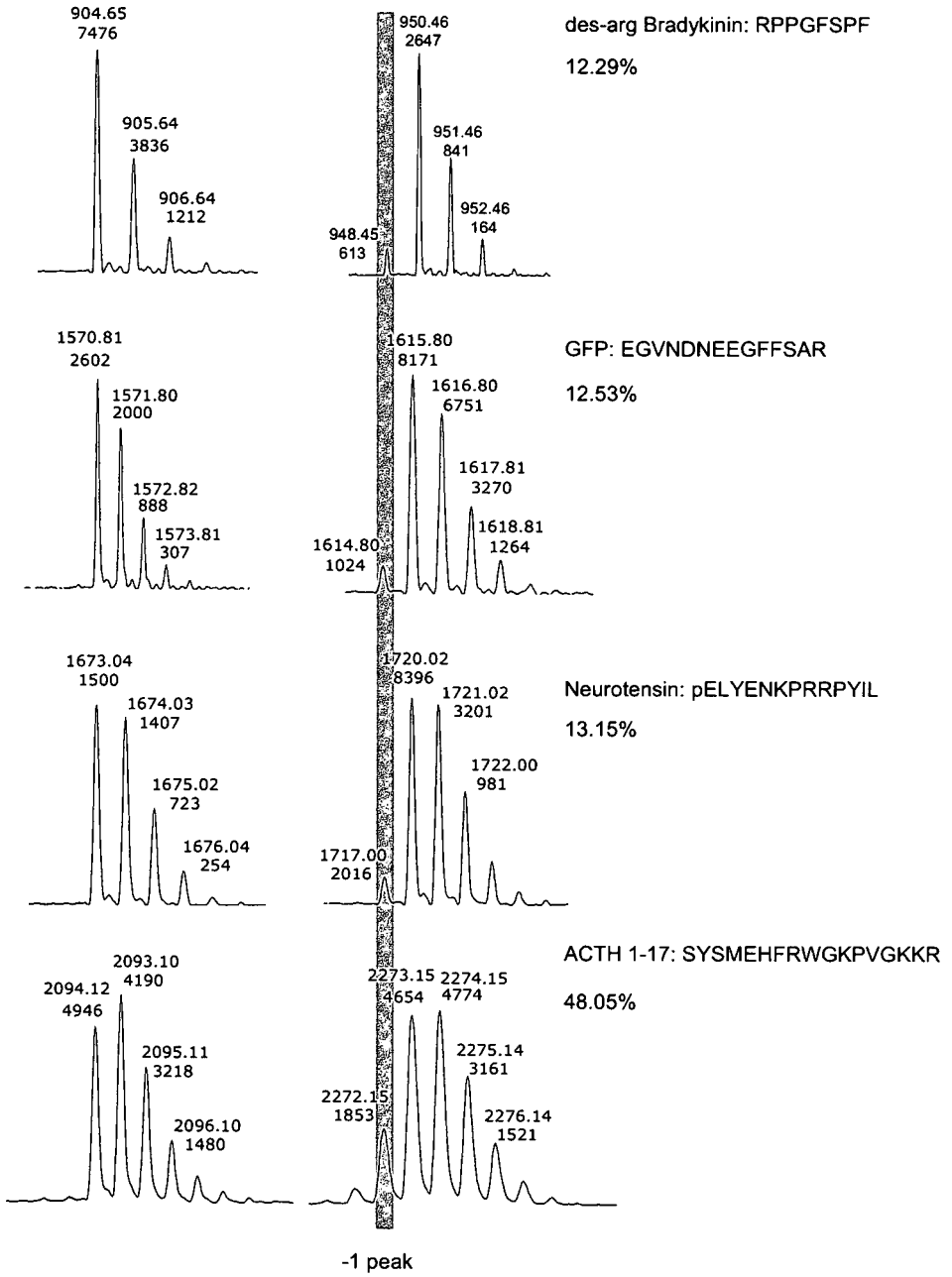


Figure 5.5. MIDAR analysis of model peptides.
A variety of model peptides were acetylated using combinatorially-coded acetic anhydride (**P**(**O**=3)=0.9 and **P**(**O**=2)=0.1) and the modified peptides were analysed by MALDI-ToF MS. The mass spectrum of the product is displayed alongside the mass spectrum of the unmodified peptide. For each, the intensity of the "minus 1" ion (shaded) was calculated relative to the monoisotopic intensity.

(a) Pyruvate kinase

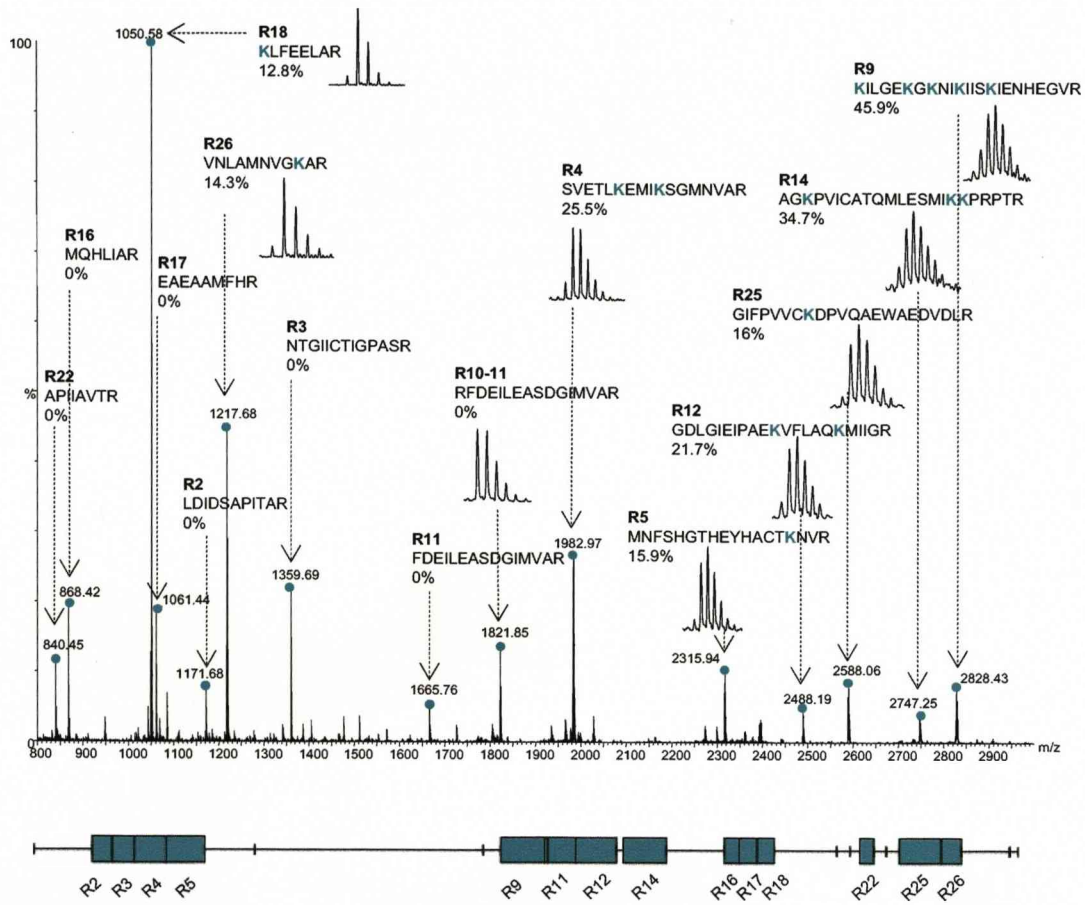
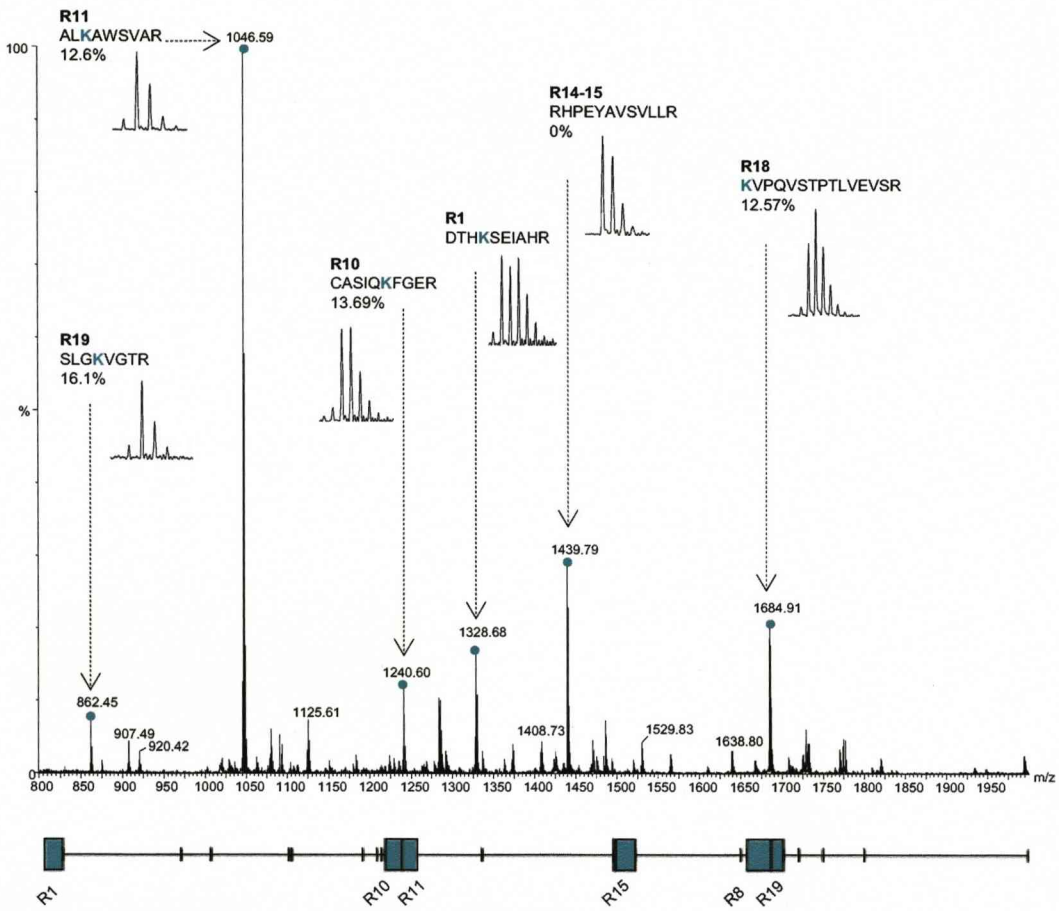
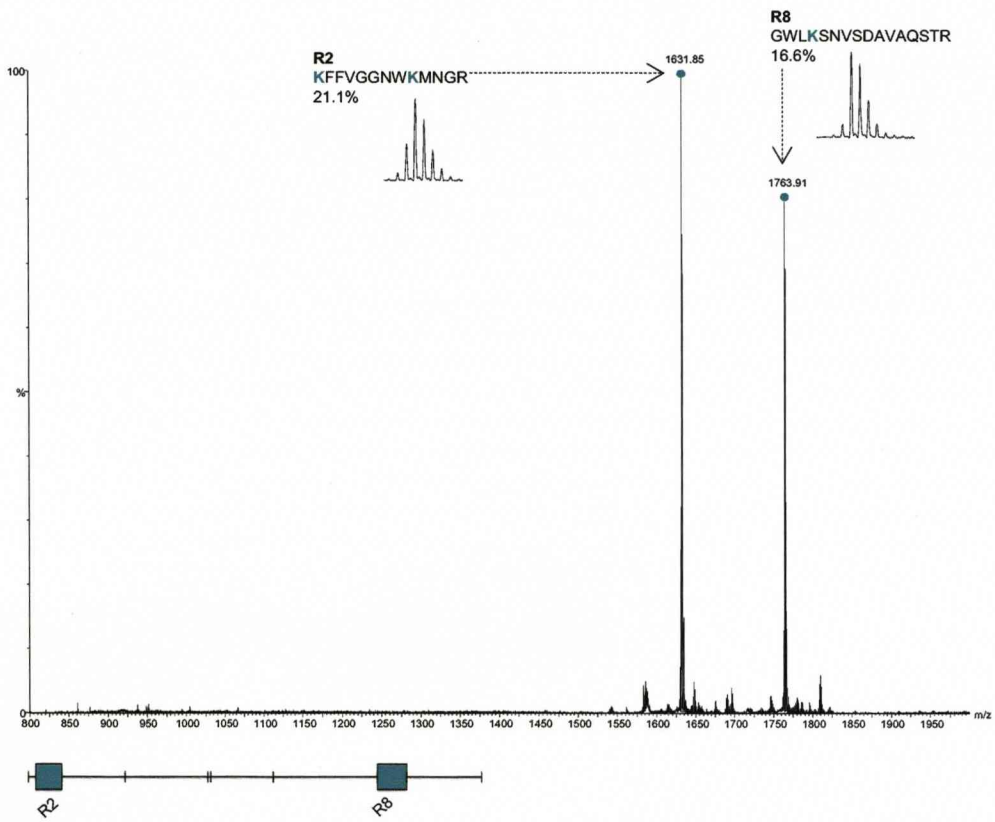


Figure 5.6. In-solution tryptic digest of MIDAR treated model proteins.
A series of purified proteins (pyruvate kinase (a), BSA (b) and triose phosphate isomerase (c)) were fully acetylated using the MIDAR reagent. The proteins were then subjected to in-solution proteolysis with trypsin, which cleaves the acetylated proteins exclusively after arginine residues. The digested peptides were analysed by MALDI-ToF MS. Each spectrum was inspected manually and the intensity of the "minus 1" ion was assessed relative to the monoisotopic ion for as many peptides as possible.

(b) BSA


R11
ALKAWSVAR
12.6%
1046.59
R14-15
RHPEYAVSVLLR
0%
R18
KVPQVSTPTLVEVSR
12.57%
R1
DTHKSEIAHR
R10
CASIQKFGER
13.69%
R19
SLGKVGTR
16.1%
100
%
0
1439.79
1684.91
1328.68
1240.60
862.45
1125.61
907.49
920.42
1408.73
1529.83
1638.80
800 850 900 950 1000 1050 1100 1150 1200 1250 1300 1350 1400 1450 1500 1550 1600 1650 1700 1750 1800 1850 1900 1950
m/z
R1
R10
R11
R15
R8
R19

(c) Triose phosphate isomerase



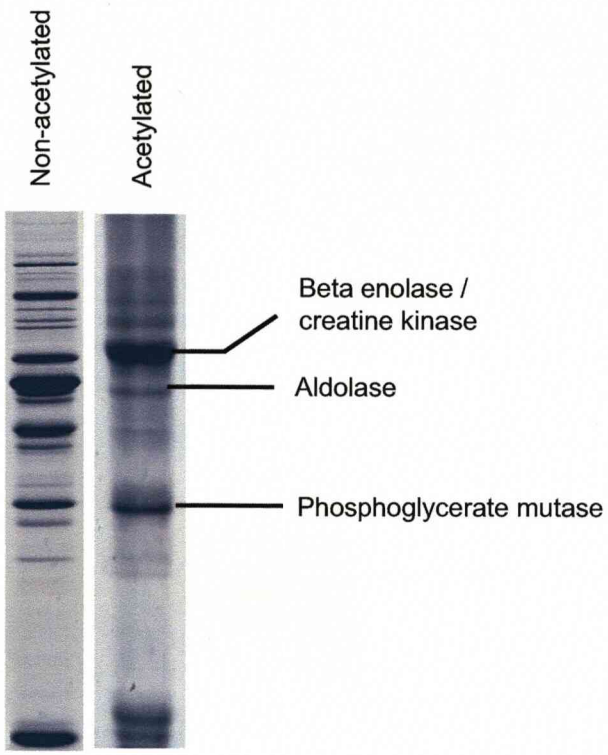


Figure 5.7. 1-D SDS-PAGE of acetylated mouse skeletal muscle soluble proteins. Mouse skeletal muscle soluble fraction (10µg) was acetylated with the MIDAR reagent. The sample was precipitated with 50µl TCA and washed with ether to remove residual acid. The pellet was resuspended in 20mM Na_2CO_3 (10µl) and separated by SDS-PAGE alongside 10µl of unmodified sample.

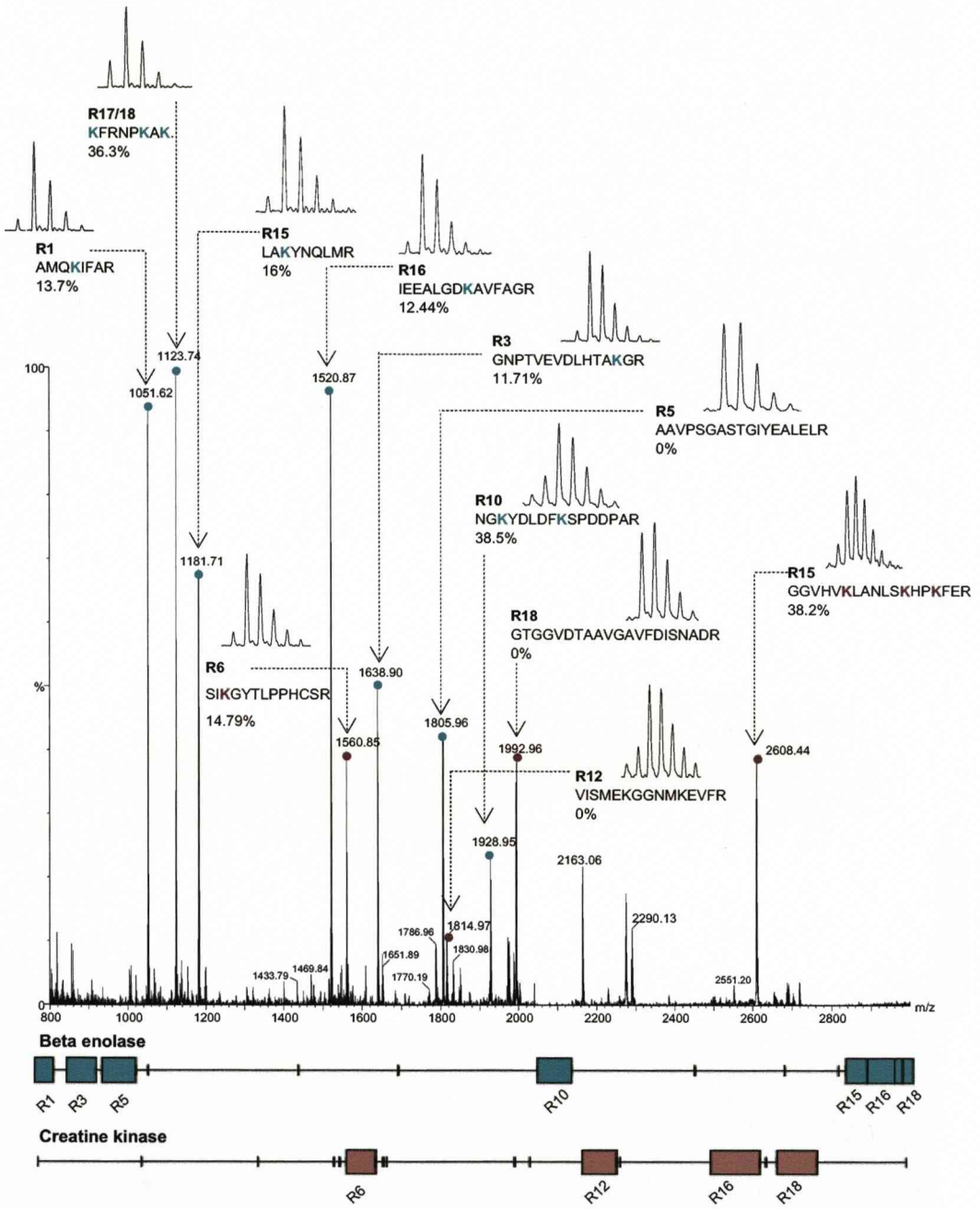


Figure 5.8. In-gel tryptic digest of MIDAR reacted mouse skeletal muscle proteins. The gel plug was excised from an SDS-PAGE gel of MIDAR reacted mouse skeletal muscle soluble proteins. The protein was digested with trypsin (150 enzyme to substrate ratio) overnight. The peptides were analysed by MALDI-ToF MS. For each peptide, the intensity of the 'minus 1" ion was calcul ated relative to the monoisotopic ion.

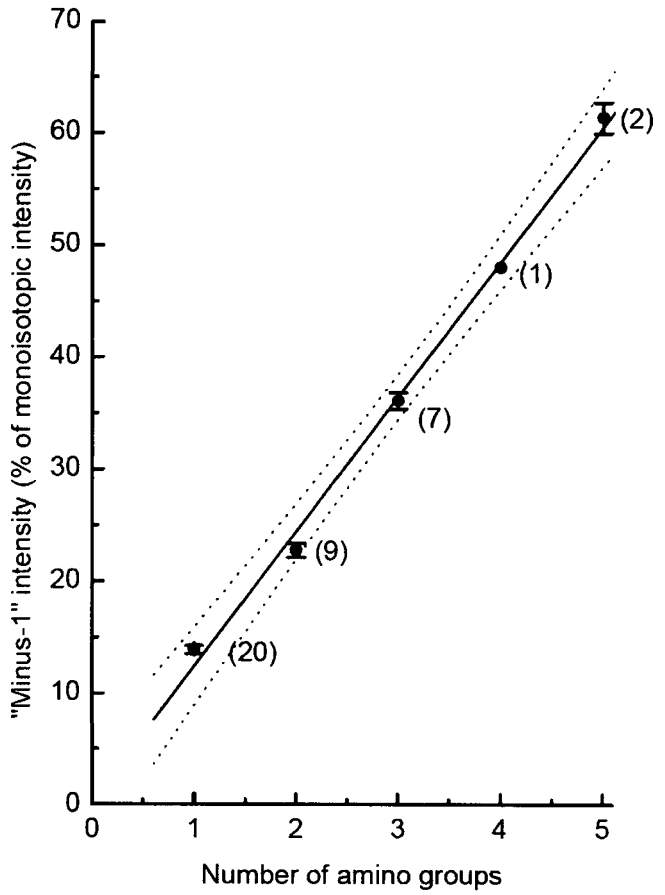


Figure 5.9. Relationship between MIDAR "'minus 1" and number of amino groups for model peptides and proteins.
A group of model peptides were fully acetylated using combinatorially-coded acetic anhydride (**P**(**O**=3)=0.9 and **P**(**O**=2)=0.1). Additionally, a series of purified proteins (BSA, TPI and PK) were acetylated prior to proteolysis with the same reagent. Finally, a natural protein mixture (soluble proteins of mouse skeletal muscle) was acetylated using the same reagent *in vitro* prior to SDS-PAGE. For each peptide, the intensity of the "minus 1" ion was calculated relative to the monoisotopic ion, and plotted according to number of free amino groups. Data are presented as mean ± SEM, for differing numbers of peptides (defined in parentheses adjacent to each symbol).

impurity that stimulated the design of the reagent described here (see Figure 5.2). The 95% confidence limits of the fitted line confirm the consistency and general utility of this reagent for a simple, integral assessment of the number of available amino groups.

5.3.3 N-terminal purification of *E. coli* proteins using the MIDAR reagent

The N-terminal positional proteomics strategy, described in Chapter 3, has considerable appeal for large-scale proteome analysis, reducing analyte complexity to one peptide per protein. The positional fixation of the peptide to the N-terminus greatly enhances search specificity and preliminary data suggest that for a preparation of N-terminal peptides (NTPeps) from a simple organism such as *E. coli*, the mass of the peptide at the mass accuracy attainable with an instrument such as an Orbitrap might be enough for unambiguous identification. Since the positional proteomic method includes an N-acetylation step, simple replacement of unlabelled acetic anhydride by the MIDAR reagent described here would simultaneously identify peptides that were naturally acetylated at the N-terminus and permit accurate (integral) determination of the number of free amino groups in each peptide.

NTPeps were prepared from exponentially grown cells of *E. coli*, using the MIDAR acetylation reagent. This peptide mixture was fractionated over a 120min RP gradient and analysed by ESI-MS using the Orbitrap mass spectrometer. For model peptides, extracted ion chromatograms for the monoisotopic mass of heavy variants were prepared in order to locate the specific MS ion. Subsequently, the mass spectra were summed across the peak and the composite mass spectrum analysed to yield the ratio of the "minus 1" ion relative to the monoisotopic ion. Representative data for two such peptides are presented in Figure 5.10. In both instances, the intensity of the "minus 1" ion (22.2% and 33.3% respectively) gives a clear indication of the number of modified amino groups (two and three respectively). Indeed, the linearity of the data obtained with the Orbitrap instrument is superior, which may reflect the markedly diminished chemical noise signal and the enhanced baseline processing/noise rejection of the Orbitrap instrument. To test the approach further, a large number of NTPeps from the *E. coli* K12 preparation were evaluated and the relative intensity of the "minus 1" ion was assessed manually (Figure 5.11). Over the mass range 600Da to 3000Da, data for over 200 peptides were manually extracted and the "minus1" intensity assessed, without knowledge of the identity of the peptide. The distribution of these values was restricted to discrete regions, and despite scatter in the data, the mean "minus 1" values (displayed on the figure with boundaries of ± 3 SD) showed no overlap, and the separation between categories

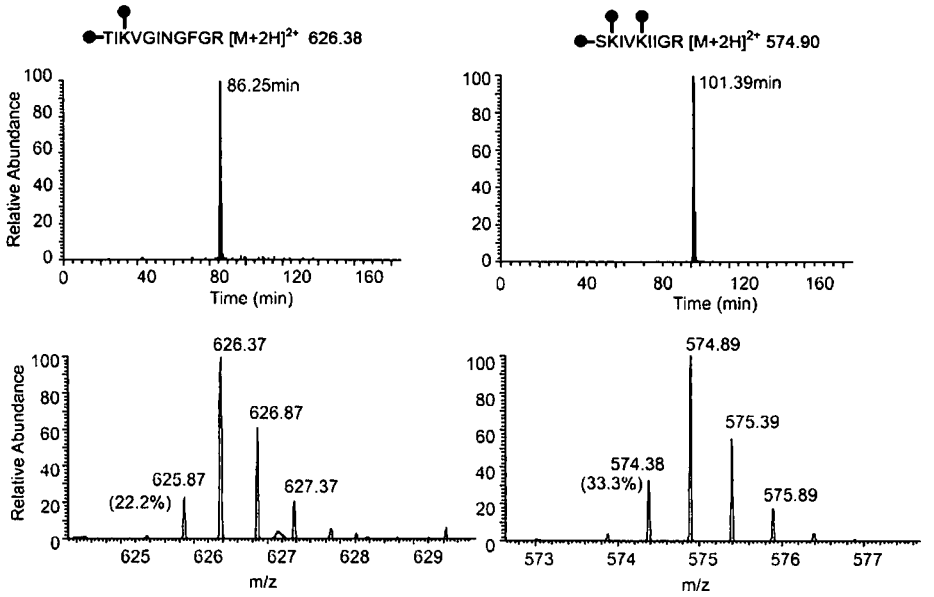


Figure 5.10. MIDAR analysis of *E.coli* N-terminal peptides.
An N-terminal peptide preparation derived from *E.coli* and grown in normal media was treated with the MIDAR reagent, and the mixture was resolved by RP chromatography and analysed by high resolution ion trap mass spectrometry on the Orbitrap instrument. Representative data are shown for two peptides. The top panels are the extracted ion chromatograms for two peptides (TIKVGINGFGR: glyceraldehyde-3-phosphate dehydrogenase A, P0A9B4 and SKIVKIIGR:enolase, P0A6P9). The lower panels are the mass spectra corresponding to those ions, and in both instances, the 'minus one'ion is clearly visible.

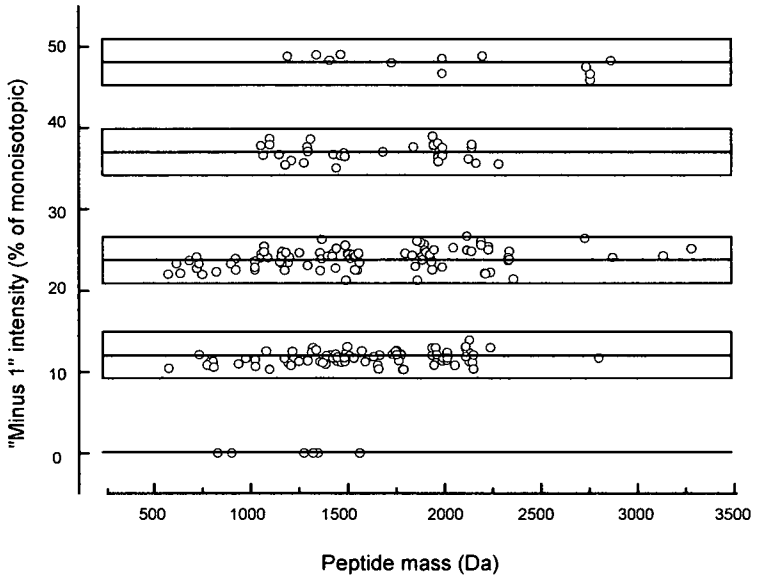


Figure 5.11. Determination of amino group frequency in *E. coli* peptides.
An N-terminal peptide preparation derived from *E.coli* and grown in normal media was treated with the MIDAR reagent, and the mixture was resolved by RP chromatography and analysed by high resolution ion trap mass spectrometry. The entire chromatogram was inspected manually, and where ions were readily identified, the intensity of the "minus 1" ion was assessed relative to the monoisotopic ion and plotted as a function of peptide mass. For each of the discrete planes of data, the error boundaries (± 3 SD, 99.7% of observations) are plotted around the mean value, to emphasis the lack of overlap.

was sufficient that an assignment of the number of amino groups would be unambiguous. For one to four amino groups, the mean "minus 1" value ± SEM were 11.6%±0.09 (n=82), 23.9%±0.12 (n=94), 37.0%±0.17 (n=34) and 48.0%±0.30 (n=12).

Part of the ambiguity in interpretation of these spectra derives from the complex overlapping sets of naturally occurring isotope distributions for the different labelled variants. If the mass spectra could be further simplified, a combination of accurate mass, MIDAR and retention time might offer a novel approach to complete proteome profiling that would lend itself to rapid genome annotation, relative or absolute quantification. To this end, *E. coli* K12 cells were grown in an isotope-depleted medium from which most of the [^{13}C] and [^{15}N] had been removed. As anticipated, the peptides derived from cells grown in this medium yielded remarkably different mass spectra (Figure 5.12) but the "minus 1" ion was clearly defined, and the intensity of this ion was slightly higher than from normal medium-grown cells, as anticipated. Again, there was a strong correlation between the number of available amino groups and the relative intensity of the "minus 1" ion relative to the monoisotopic ion (data not shown, regression of % 'minus 1" ion intensity on number of amino groups: slope = 12.55% ± 0.23%, intercept = -1 ± 0.4, correlation coefficient = 0.997, n=17, $p<0.001$). Coding of a simple acetylation reagent thus gives unequivocal encoding of the integral number of such reactive groups within a peptide, and has general utility in peptide characterisation and even, identification of post-translational modifications that are targeted to lysine residues.

The set of N-terminal peptides manually analysed from the *E. coli* MIDAR N-terminal preparation, were also analysed in terms of their retention time (Figure 5.13). As expected, there was some correlation between retention time and mass of the peptides, suggesting that these are not two entirely independent parameters for proteome characterisation. Furthermore, the predicted set of N-terminal peptides from *E. coli* K12 were analysed according to mass, predicted retention time (Krokhin *et al.*, 2004) and number of lysine residues between the most accessible masses of 1000Da and 5000Da (Figure 5.14). Again, there was correlation between retention time and mass of the peptides. The abundance of lysine residues in most proteomes is around 5% (MSDB ver 08/09/2006, lysine frequency = 5.4%). Knowledge of the number of amino groups (lysine residues with or without an exposed α-amino group) in a peptide therefore has the potential to increase the efficiency of database searching. For example, two N-terminal peptides (YFAU_ECOLI: MNALLSNPFKER and MALG_ECOLI: MAMVQPKSQKAR) have masses of 1463.734 and 1463.738Da in their N-acetylated forms respectively, a separation of approximately 3ppm. That difference would be

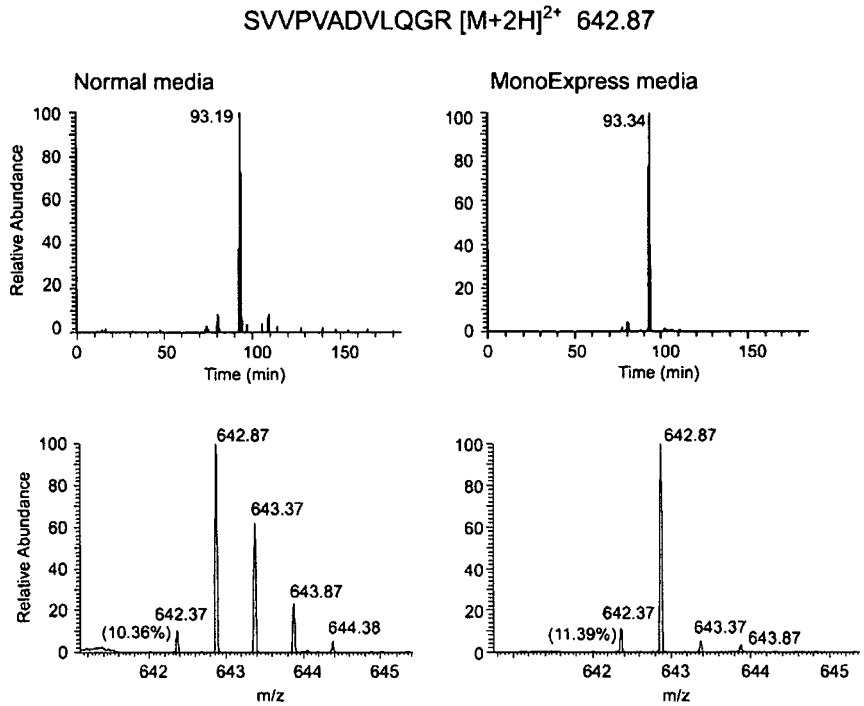


Figure 5.12. MIDAR analysis of N-terminal peptides from *E. coli* grown in isotopically depleted media.
Representative data are shown for a peptide (SVVPVADVLQGR, asparaginyl-tRNA synthetase, P0A8M0) from cells grown in normal media and isotopically depleted media. The top panels are the extracted ion chromatograms for the peptide grown in either media. The lower panels are the mass spectra corresponding to those ions, and in both instances, the 'minus one"ion is clearly visible.

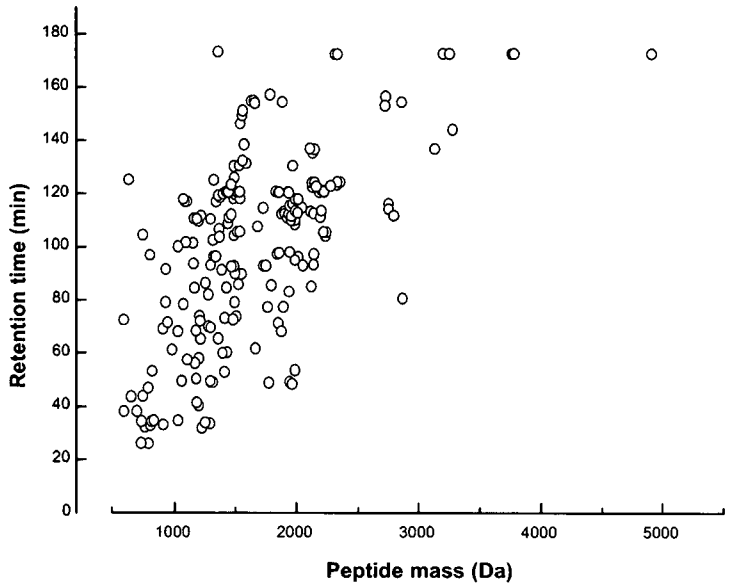


Figure 5.13. Relationship between mass and retention time for *E. coli* N-terminal peptides.
The N-terminal peptide preparation derived from *E.coli* and grown in normal media was treated with the MIDAR reagent. The mixture was resolved by RP chromatography and analysed by high resolution ion trap mass spectrometry. The entire chromatogram was inspected manually, and where ions were readily identified, the retention time was determined and plotted as a function of peptide mass.

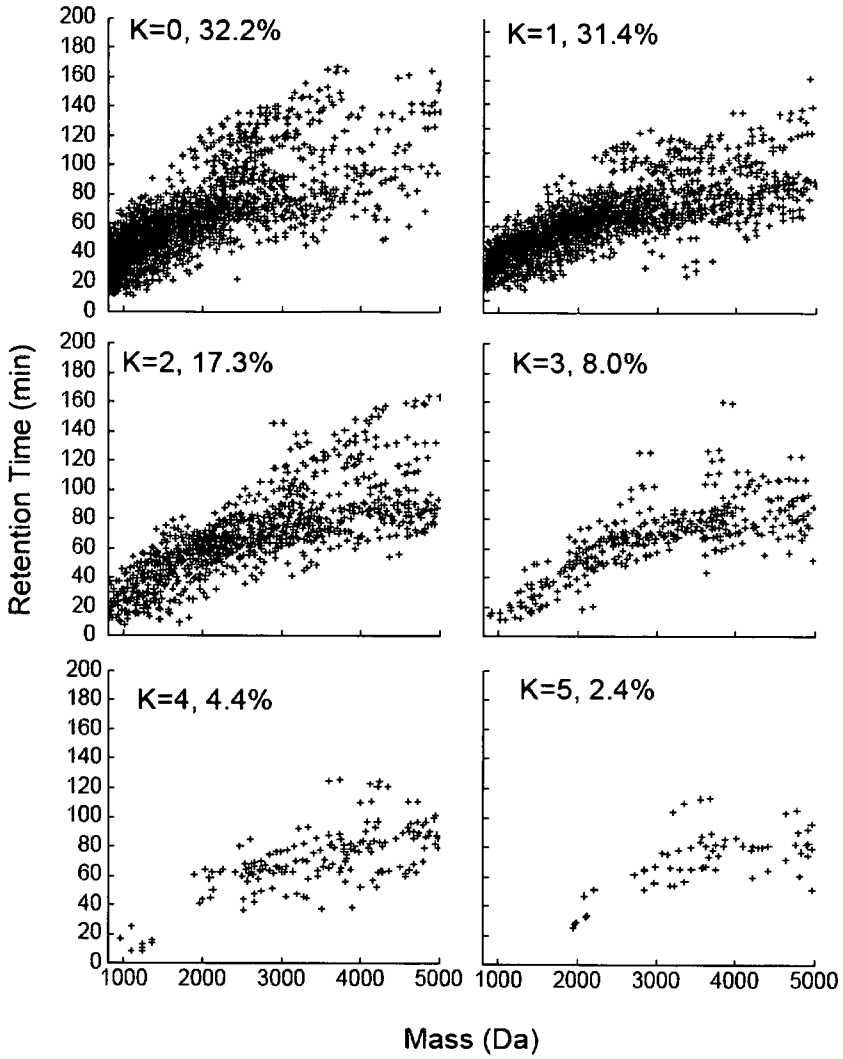


Figure 5.14. Theoretical analysis of *E.coli* N-terminal peptides.
A database of predicted N-terminal peptides from *E.coli* K12 was compiled and used to generate a data set of mass, chromatographic retention time and number of lysine residues. The calculation of retention time was not derived according to a specific chromatographic system but serves to distribute the peptides in chromatographic space, using the default parameters embedded in the software (A=8, B=1.5). Furthermore, the addition of acetyl groups would be expected to modify the retention time, probably delaying peptide elution to a small degree. The data set was distributed over the mass, retention time axes and is presented for peptides containing between zero and five lysine residues.

difficult to resolve on mass alone but because the first peptide contains a single lysine residue and the latter contains two, they occupy discretely resolved lysine 'planes' and are readily resolved, permitting unambiguous identification of both spectra. For peptides larger than 1000Da at 20ppm mass resolution, there are 1132 instances of peptides having unresolvable mass, and of these, 699 (62%) could be resolved in the lysine plane. At 10ppm, 5ppm and 2ppm, knowledge of the lysine count resolved 61%, 57% and 39% of the contentious peptides, respectively. Most (80%) of NTpeps were characterised as containing zero, one or two lysine residues, yielding "minus one" intensities between 0 and 30%, depending on the acetylation state of the true N-terminus. This might argue a case for increasing the proportion of the "minus one" component to greater than 10%, which would increase the separation between peptides that contain the most common numbers of lysine residues, although our analysis of over 200 NTpeps suggests this is not necessary. A strategy based on positional proteomics, isotopically-depleted media, and MIDAR could generate a substantially simplified analyte mixture that would allow extensive profiling of a proteome in a single experiment, permitting higher levels of biological replication and increased statistical certainty.

There are other advantages to this approach. First, because the preparation of positionally located peptides is essentially a self-cleaning method (McDonald *et al.,* 2005; McDonald and Beynon, 2006); partially acetylated peptides would be removed by the clean up steps. An added advantage of lysine modification is that it removes ambiguity between lysine and glutamine residues, which differ in mass by 0.036Da.

5.3.4 Effect of acetylation on retention time

It is possible to predict the retention time of a peptide by virtue of its amino acid composition and the conditions required for chromatographic separation (slope of gradient and delay time; Krokhin *et al.,* 2004). However, modification of a peptide can affect the rate at which it is eluted from the RP media, at a rate dependent on the nature of the modification. For example, the small hydrophilic $-PO_3H_2$ group has only a minor affect on the chromatographic retention time of peptides (Browne *et al.*, 1982) To be able to effectively predict the retention time of acetylated NTpeps it is necessary to determine the extent to which acetylation effects the retention time of peptides. For this purpose, two purified proteins (bovine serum albumin (BSA) and pyruvate kinase (PK)) were digested with trypsin and differentially modified with acetic anhydride prior to chromatographic fractionation.

Calibration

Retention time (RT) vs. hydrophobicity (HP) of Peptides, is a linear function (Krokhin *et al.*, 2004):

$$RT=A+B*(HP);$$

Where intercept A is the gradient delay time (individual for each HPLC system used) and slope B is a value related to the slope of acetonitrile gradient. B is constant for different HPLC systems as long as the same slope of the linear gradient is used. Before investigating the effect of acetylation on retention time it was necessary to determine the slope of the acetonitrile gradient and the gradient delay time (intercept). This was achieved using a tryptic digest of purified BSA. The digested peptides were analysed by LC-MS/MS on the ion trap instrument using the extended, three hour RP gradient. The retention times of the matched peptides were determined by preparing extracted ion chromatograms on the MS/MS data. Hydrophobicities for identified peptides were calculated using the Sequence Specific Retention (SSR) Calculator (Table 5.1; Krokhin *et al.*, 2004). RT was plotted against relative HP for each BSA peptide (intercept = 2.97 (A), slope = 1.60 (B); Figure 5.15). The SSR calculator was then used to determine the theoretical retention times for the BSA peptides analysed and the results were plotted alongside the experimental data (Figure 5.16a). The change in RT (ΔRT) between the experimental and theoretical times were also plotted (Figure 5.16b). This data suggests a loose correlation between ΔRT and HP.

Retention time of differentially acetylated purified protein digests

To ascertain the affect of acetylation on peptide retention time, two purified proteins were proteolysed with trypsin and partially acetylated, to ensure that both unmodified and acetylated intermediates were present. The tryptic digest was split into two aliquots, the first was treated with 0.1mg sulfo-NHS acetate in (dissolved in 20mM Na_2C0_3 pH 8.5), for 10 min (for incomplete acetylation) and the second remained unmodified. The acetylation reaction was quenched by the addition of 5mg polymer bound Tris, which was subsequently removed by centrifugation. The two aliquots (modified and unmodified), were combined, diluted 1 in 50 using 0.1% (v/v) FA and analysed by LC-MS/MS on the LTQ ion trap instrument, using a three hour RP gradient. The MS/MS data were used to search the SwissProt database using the MASCOT search engine.

Retention times were determined by preparing extracted ion chromatograms for identified peptides from the MASCOT search (Figure 5.17). The theoretical retention times were predicted using the SSR calculator for comparison (Table 5.2). Retention times for

Peptide	Sequence	HP	RT
4	DLGEEHFK	16.23	21.21
6	LVNELTEFAK	27.42	49.89
10	ETYGDMADCCEK	11.81	21.21
11/12	QEPERNECFLSHK	14.71	18.47
12/13	NECFLSHKDDSPDLPK	22.09	32.76
14	LKPDPNTLCDEFK	22.92	43.45
18	YLYEIAR	25.97	37.88
21	YNGVFQECCQAEDK	18.19	32.43
22	GACLLPK	18.33	25.26
31	AWSVAR	15.88	22.58
35	LVTDLTK	16.36	24.93
37	ECCHGDLLECADDR	16.63	28.99
37/38	ECCHGDLLECADDRADLAK	22.48	37.72
39	YICDNQDTISSK	13.12	21.95
40/41	LKECCDKPLLEK	16.39	18.04
50	EYEATLEECCAK	16.4	28.81
51/52	DDPHACYSTVFDKLK	23.36	47.09
53	HLVDEPQNLIK	22.12	36.81
57/58	KVPQVSTPTLVEVSR	23.15	36.81
58	VPQVSTPTLVEVSR	23.92	41.81
67	RPCFSALTPDETYVPK	27.66	47.46
72	QTALVELLK	30.68	55.28
78	LVVSTQTALA	23.42	45.17

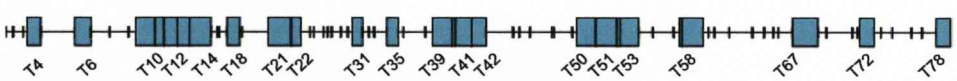


Table 5.1. Relative hydrophobicity and retention time of BSA.
BSA (10µg, Pierce) was reduced and alkylated prior to digestion with trypsin (1:50 enzyme: substrate ratio). The digested peptides were analysed using a three hour RP gradient on the ion trap instrument. MS/MS data were used to search the SwissProt database using the MASCOT search engine. The taxonomy was restricted to *Homo sapiens;* fixed modifications: carbamidomethylation of cysteine; variable modification: oxidation of methionine; protease: trypsin; missed cleavages: 1; peptide tolerance: 1.5Da; MS/MS tolerance: 0.6Da; instrument: ESI-TRAP; peptide charge: 1+, 2+ and 3+. Protein identifications with a Mowse score greater than 50 were accepted as confident identifications. Extracted ion chromatograms were prepared for BSA peptides using the values matched in MASCOT and retention times (RT) recorded. Relative hydrophobicity (HP) was determined from the amino acid sequences using the sequence specific retention calculator.

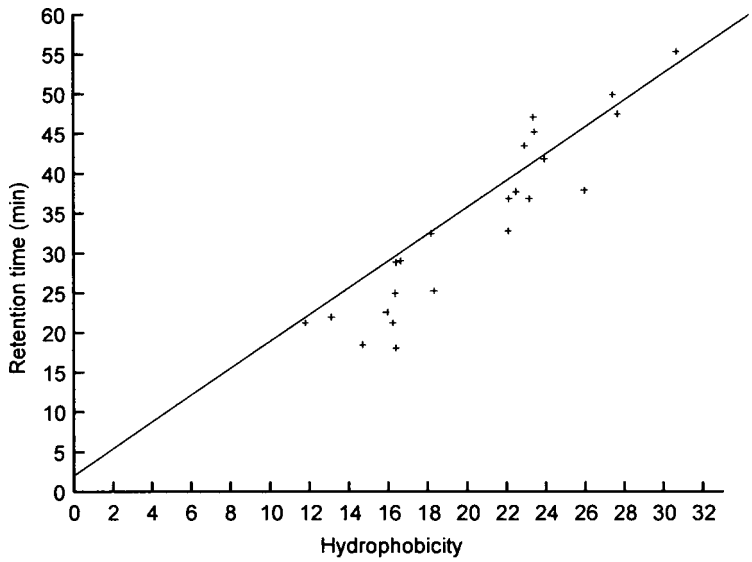


Figure 5.15. Calibration of chromatography gradient using BSA.
Purified BSA (10μg, Pierce) was reduced and alkylated prior to digestion with trypsin (1:50 enzyme: substrate ratio). The digested peptides were analysed using a three hour RP gradient on the ion trap instrument. MS/MS data were used to search the SwissProt database using the MASCOT search engine. The taxonomy was restricted to mammalia; fixed modifications: carbamidomethylation of cysteine; variable modification: oxidation of methionine; protease: trypsin; missed cleavages: 1; peptide tolerance: 1.5Da; MS/MS tolerance: 0.6Da; instrument: ESI-TRAP; peptide charge: 1+, 2+ and 3+. Extracted ion chromatograms were prepared using the values of peptides matched in MASCOT and retention times recorded. Relative hydrophobicity (determined from the amino acid sequences using the SSR calculator) was plotted against the experimental RT.

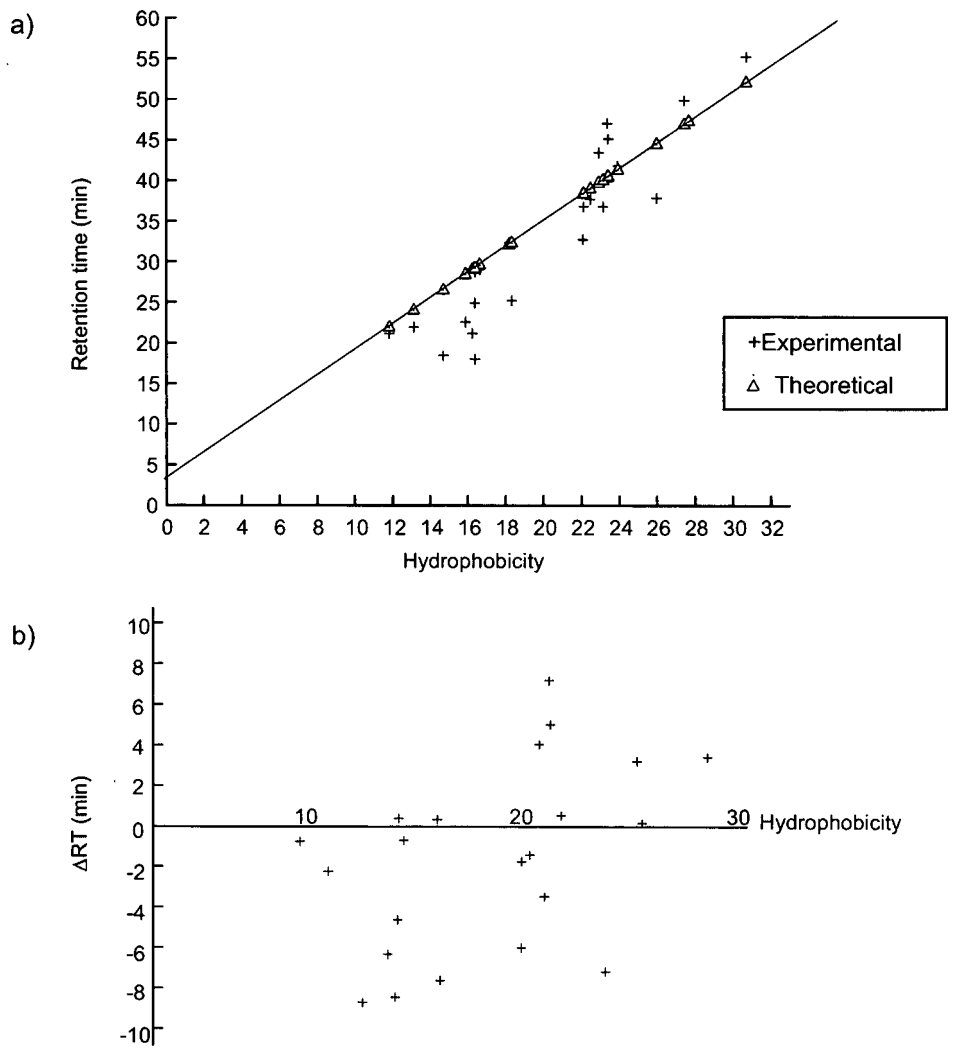


Figure 5.16. Comparison of experimental and theoretical retention times for BSA peptides.
The values obtained from the slope and the intercept of the plot (experimental data) were used to calculate the theoretical RT from the same group of peptides (SSR calculator) and the data plotted onto the same graph (a). The Δ RT between the experimental RT and theoretical RT was also plotted (b).

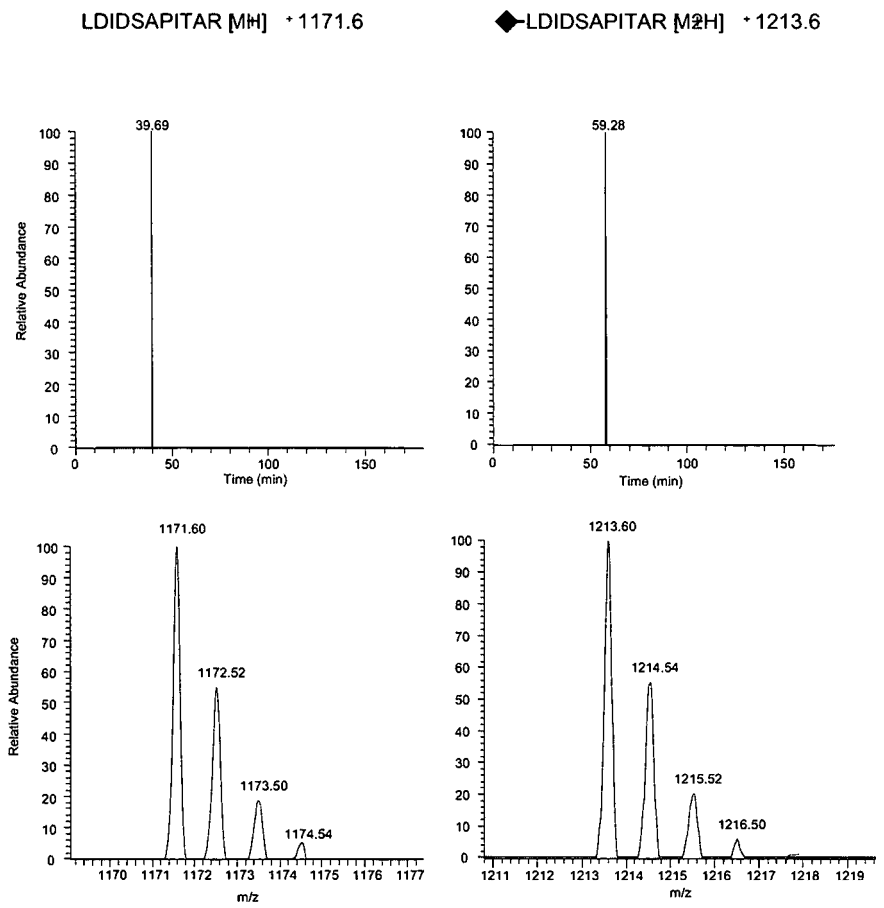


Figure 5.17. Effect of acetylation on peptide retention time.
Purified proteins (BSA and PK) were proteolysed by trypsin prior to incomplete acetylation with sulfo-NHS acetate. The peptide mixtures were analysed using the LTQ ion trap instrument. Extracted ion chromatograms were prepared for non-acetylated and acetylated forms of each peptide. Representative data are shown for (LDIDSAPITAR from PK). The top panels are extracted ion chromatograms and the lower panels are the mass spectra corresponding to those ions.

unmodified and differentially acetylated forms of the peptides, along with the theoretical retention times (RT_{th}), were plotted (Figure 5.18). Tryptic peptides were plotted according to increasing $[M+H]^+$. In general, the theoretical and observed retention times are comparable and correlate with peptide size. However, in the case of BSA, the two largest peptides (11/12 and 37/38) display relatively short retention times in relation to the other peptides. Both of these peptides contain internal arginines along with a high content of other hydrophilic residues (glutamate, aspartate and histidine), which may cause them to elute from the RP media early. In cases for which there are two possible acetylation sites on the peptide it is apparent that the second acetyl group causes a smaller ΔRT than addition of the first. One possible explanation for this would be that the initial acetyl group is being placed on the ε-amino group causing the otherwise hydrophilic lysine residue to become substantially more hydrophobic thus displaying a marked increase in retention. The second acetyl group, which will be located on the α-amino group, will cause a smaller ΔRT compared with the first modification. This theory is supported by the difference in pKa exhibited by α and ε-amino groups (9.04 and 10.4 respectively; Mikami and Takao, 2007). Since the acetylation reaction is performed at high pH (0.5M Na_2CO_3), it is likely that the ε-amino groups will be more reactive under these conditions.

Retention time of *E. coli* NTpeps

The LC-MS/MS gradient used for fractionation of *E. coli* NTpeps prior to analysis on the Orbitrap mass spectrometer, was much shallower than that used for the ion trap instrument (slope 2.48 compared to 1.60). For this reason the affect of modification on peptide retention time will be much more apparent. To determine the extent of ΔRT for the acetylated *E. coli* NTpeps, a range of known peptides were used to compare theoretical RT values, for unmodified peptides, with observed values from the acetylated peptides. The values obtained for a set of 25 known *E. coli* NTpeps are represented in Table 5.3. In each case ΔRT was used in combination with the number of acetylation sites (n) to calculate the shift in retention time for one acetyl group. The resulting values were highly variable, ranging from 11.37 min to 62.32 min. This wide degree of variation in retention time will make the prediction of acetylated retention times using existing algorithms highly challenging. For this reason it is unlikely that the AMT strategy, at present, will be successful when applied to the analysis of NTpeps.

BSA

Peptide	Sequence	$[M+H]^+$	No. amino groups	Predicted RT	Observed RT		
					0	1	2
31	AWSVAR	689.37	1	27	22.58	41.58	
22	GACLLPK	759.4	2	31.2	25.26	45.07	63.7
35	LVTDLTK	789.46	2	27.8	24.93	56.14	75.12
18	YLYEIAR	927.48	1	44.1	37.88	66.07	
4	DLGEEHFK	974.45	2	27.6	21.21	32.28	49.67
78	LVVSTQTALA	1002.58	1	39.8	45.17	64.31	
72	QTALVELLK	1014.61	1	52.2	55.2	73.4	
6	LVNELTEFAK	1163.62	2	46.6	49.89	82.49	98.31
58	VPQVSTPTLVEVSR	1511.84	2	40.7	41.8	76.52	93.83
11/12	QEPERNECFLSHK	1674.75	2	25	18.47	27.82	40.46
37/38	ECCHGDLLECADDRADLAK	2250.89	2	38.2	37.72	44.77	53.97

PK

Peptide	Sequence	$[M+H]^+$	No. amino groups	Predicted RT	Observed RT		
					0 Ac	1 Ac	2 Ac
25	DIQDLK	731.39	2	21.8	16.24	36.91	60.21
53	APIIAVTR	840.52	1	32.2	27	53.81	
45	MQHLIAR	868.47	1	25.4	13.78	37.2	
44	GDYPLEAVR	1019.51	1	37.9	34.5	50	
46	EAEAAMFHR	1061.48	1	26.8	19.8	38.23	
3	LDIDSAPITAR	1171.62	1	38.2	39.69	59.28	
4	NTGIICTIGPASR	1302.7	1	41.7	37.73	52.46	
20	VYVDDGLISLQVK	1448.79	2	53.9	59.98		95.13
35/36	RFDEILEASDGIMVAR	1821.91	1	57.2	61.3	76.83	
8	MNFSHGTHEYHAETIK	1901.85	2	33.2	18.7		40.45
10	TATESFASDPILYRPVAVALDTK	2465.29	2	57.8	67.31		87.24
43	AEGSDVANAVLDGADCIMLSGETAK	2437.12	2	52.6	76.5		99.43

Table 5.2 Effect of acetylation on peptide retention time on peptides from purified proteins.

Purified BSA and PK (10µg, Pierce) were reduced and alkylated prior to digestion with trypsin (1:50 enzyme: substrate ratio). The digested peptides were incompletely acetylated to yield a mixture of unmodified, partially acetylated and fully acetylated peptides. The peptide mixtures were fractionated using a three hour RP gradient and analysed by LC-MS/MS on the LTQ ion trap instrument. MS/MS data was used to interrogate the SwissProt database using the MASCOT search engine. The taxonomy was restricted to mammalia; fixed modifications: carbamidomethylation of cysteine; variable modification: oxidation of methionine, acetyl lysine and acetyl N-terminal; protease: trypsin; missed cleavages: 1; peptide tolerance: 1.5Da; MS/MS tolerance: 0.6Da; instrument: ESI-TRAP; peptide charge: 1+, 2+ and 3+. Identified peptides (unmodified and acetylated versions) were used to prepare extracted ion chromatograms from the raw LC-MS/MS data in order to determine peptide retention times. Theoretical retention times were determined using the peptide sequence and the SSR calculator.

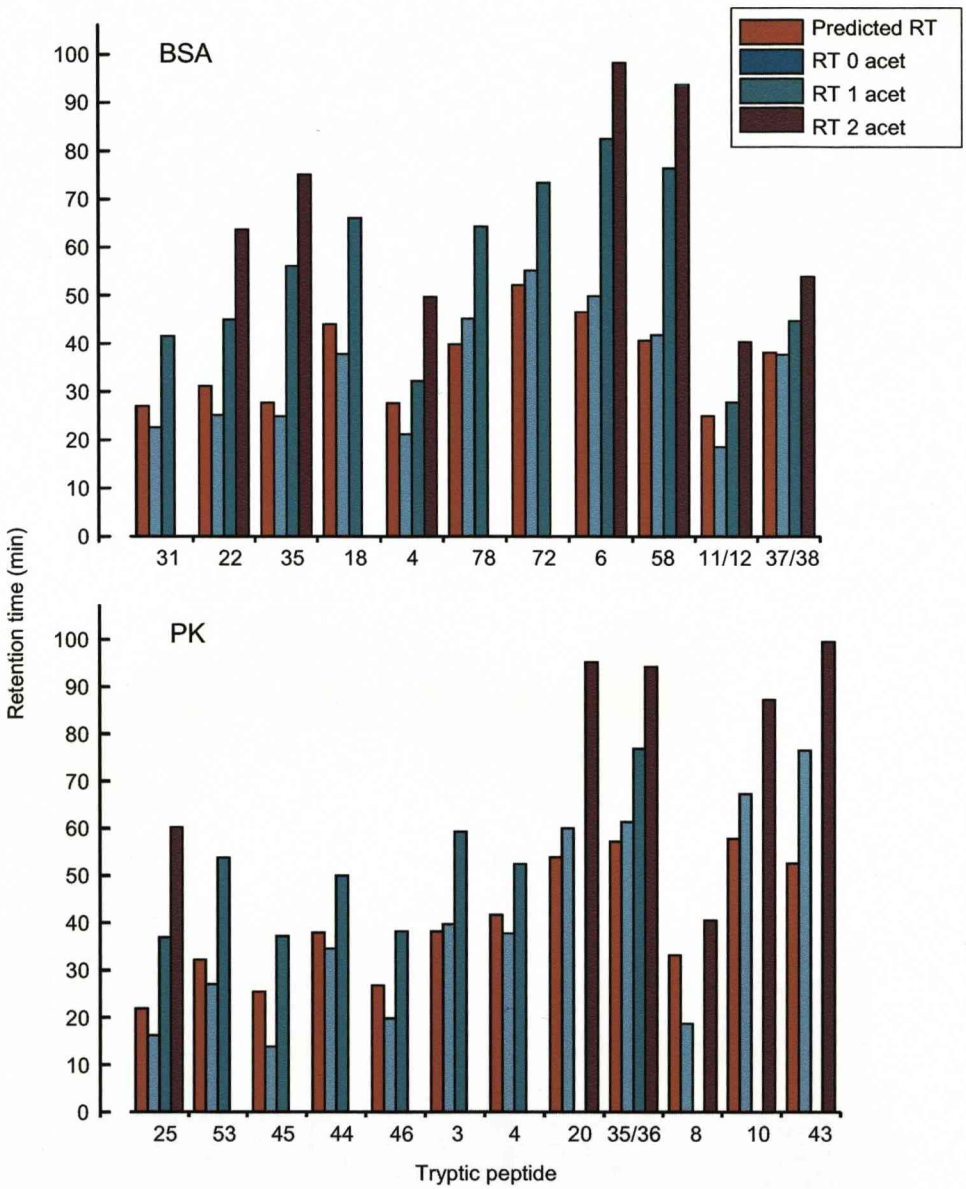


Figure 5.18. Retention time of unmodified and acetylated peptides from purified proteins.
The observed retention times were plotted alongside the theoretical retention times (determined using the peptide seqence and t he SSR calculator) in order of increasing peptide mass.

Protein (Ac)	$[M+H]^+$	Sequence	No. acet groups (n)	RT_{th}	observed RT (0 acet)	Δ RT
D-erythro-7,8-dihydroneopterin triphosphate epimerase (P0AC19)	880.51	AQPAAIIR	1	38.8	52.95	14.15
Histidyl-tRNA synthetase (P60906)	996.57	AKNIQAIR	2	28.5	54.62	26.12
Elongation factor Tu (EF-Tu) (P0A6N1)	1048.52	SKEKFER	3	15.6	49.71	34.11
2,3-bisphosphoglycerate-dependent phosphoglycerate mutase (P62707)	1081.69	AVTKLVLVR	2	57.6	100.64	43.04
Protein ygiN (P0ADU2)	1086.61	MLTVIAEIR	1	70.9	128.67	57.77
Esterase yeiG (P33018)	1115.47	MEMLEEHR	1	42.9	117.2	74.3
Enolase (P0A6P9)	1138.71	SKIVKIIGR	3	45.6	101.51	55.91
Long-chain fatty acid transport protein (P10384)	1180.5	TILNHTLGFPR	1	67.8	102.64	34.84
30S ribosomal protein S20 (P0A7U9)	1182.67	ANIKSAKKR	4	-0.1	49.63	49.73
30S ribosomal protein S3 (P0A7V5)	1188.64	GQKVHPNGIR	2	9.6	40.33	30.73
10 kDa chaperonin (groES protein) (P0A6G1)	1192.61	MNIRPLHDR	1	41.8	58.03	16.23
Dihydroorotase (P05020)	1195.69	TAPSQVLKIR	2	40.7	74	33.3
Putative 6-pyruvoyl tetrahydrobiopterin synthase (P65870)	1195.69	STTLFKDFTFEAAHR	2	81.1	129	47.9
Glyceraldehyde-3-phosphate dehydrogenase A (P0A9B4)	1244.69	TIKVGINGFGR	2	54.7	86.34	31.64
Phosphoribosylaminoimidazole-succinocarboxamide synthase (P0A7D9)	1249.61	MQKQAELYR	2	38	61.66	23.66
Biotin carboxylase (P24182)	1255.7	MLDKIVIANR	2	57.3	105.36	48.06
Pantoate--beta-alanine ligase (P31663)	1352.81	MLIIETLPLLR	1	111.1	173.42	62.32
HIT-like protein ycfF (P0ACE7)	1389.75	AEETIFSKIIR	2	61.1	120.08	58.98
Elongation factor Ts (EF-Ts) (P0A6P1)	1412.79	AEITASLVKELR	2	62.8	120.87	58.07
60 kDa chaperonin (groEL protein) (P0A6F7)	1416.7	AAKDVKFGNDAR	3	18.6	60.38	41.78
Glutaminyl-tRNA synthetase (Q8FJW4)	1431.71	SEAEARPTNFIR	1	47	61.88	14.88
S-ribosylhomocysteine lyase (P45578)	1441.72	PLLDSFTVDHTR	1	67.5	110.76	43.26
Peptide deformylase (P0A6K3)	1446.78	SVLQVLHIPDER	1	69	116.37	47.37
Phosphoglycerate kinase (P0A799)	1671.89	SVIKMTDLDLAGKR	3	67.9	107.61	39.71
Beta-lactamase (P62593)	2132.3	HPETLVKVKDAEDQLQR	3	52.2	97.28	45.08

Table 5.3. Effect of acetylation on the retention time of *E. coli* N-terminal peptides.
The *E. coli* N-terminal preparation generated using the MIDAR reagent, was analysed by LC-MS on the Orbitrap instrument. Extracted ion chromatograms were prepared using m/z values from previously identified N-terminal peptides. Peptide retention times were recorded along with the theoretical values for the unmodified peptides (determined using the SSR calculator).

5.4 SUMMARY

This chapter reports a new isotope coded acetylation reagent that comprises a mixture of two differentially labelled chemicals, separated by 1Da. The reagents are mixed asymmetrically (10% of the lighter and 90% of the heavier variant). When peptides are labelled with this reagent, the isotope pattern of the products is complex, but the ratio of the two specific ions gives a precise measure of the number of amino groups in each peptide. This approach, termed MIDAR, may be of value in global proteomics, as it resolves a peptide mixture into a new fundamental set, reflecting the number of amino groups in the peptide. The approach is generally applicable to any peptide-based proteome analysis, but is illustrated here by a strategy of positional proteomics based on the analysis of N-terminal peptides. Knowledge of the number of amino groups can provide an integral statistic that can be used as an added parameter in database searching strategies.

6. CONCLUSIONS

To date, post-genomic research has focused on quantifying changes in DNA and mRNA expression as a result of disease or other environmental factors (Duggan *et al.*, 1999; Quackenbush, 2006; Hoheisel, 2006). However, DNA and mRNA changes do not always correlate with changes in protein expression (Anderson and Seilhamer, 1997; Chen *et al.*, 2002; Gygi *et al.*, 1999). Since proteins are the functional representations of the genes which encode them, proteomics is rapidly becoming the method of choice for life science research.

The primary challenge in proteomics is identification, characterisation and quantification of all proteins expressed within any proteome. There is a pressing need for strategies and reagents that reduce the complexity of a total proteome analyte, particularly when the first step in proteome characterisation is the increase in complexity brought about by the total proteolytic digestion, usually with trypsin, which increases the number of analyte species by about 30-50 fold.

Traditionally, the main methods for proteome characterisation have been protein separation by 1-D and 2-D SDS-PAGE, followed by proteolysis of resolved proteins and MS analysis. More recently, there has been a move towards global methods of proteome characterisation, employing direct analysis of complex protein mixtures. However, the complete set of peptides generated from an entire proteome will potentially contain many hundreds of thousands of peptides, which generates a formidable analytical challenge. Even with the benefit of 2-D chromatography (MudPIT), it is likely that the mixture will deliver more peptides to the mass spectrometer than can feasibly be analysed in the time frame it takes the ionised sample to reach the ESI source. Furthermore, data-dependent acquisition is likely to direct MS analysis to peptides originating from the most abundant proteins in the mixture, thus limiting dynamic range.

The growing popularity of so called "shotgun" MS approaches has led to a plethora of strategies aimed at purifying a subset of target peptides in an attempt to gain simplification of a complex analyte mixture. These methods aim to eliminate the majority of the proteome-derived peptides but retain sufficient information necessary for analysis. An efficient strategy for proteome simplification is the isolation of a single peptide from every protein within the mixture. Furthermore, if the location of the peptide is anchored to a precise position within the parent protein, this will reduce the amount of search space required for identification. The

most obvious regions for positional specific isolation are the extremities of protein molecules (i.e. the N or C-termini).

6.1 N-TERMINAL POSITIONAL PROTEOMICS

The primary aim of this thesis was to develop an efficient strategy for N-terminal peptide isolation that would reduce the complexity of a complete proteome and provide added 'positional' information for database searching.

The initial N-terminal "positional proteomics" strategy developed consisted of a series of a chemical derivatisation steps, followed by affinity capture of internal peptides, as follows:

1. Acetylation of intact proteins in their native state in order to block primary amino groups (α and ε).
2. Proteolysis of acetylated proteins using trypsin to yield a mixture of N-terminally blocked (N^{α}-acetylated) N-terminal and unblocked internal, arginine terminated peptides.
3. Biotinylation of the peptide mixture, using an NHS-ester of biotin, resulting in a mixture of biotinylated internal and non-biotinylated (N^{α}-acetylated), N-terminal peptides.
4. Affinity capture of biotinylated internal peptides using streptavidin Sepharose resulting in a preparation enriched in N-terminal peptides, which can then be analysed by MS.

This first method, although effective, required multiple peptide-purification steps to separate the peptides from the excess reagents used. The consequence of this is a reduced yield of material (N-terminal peptides). The original method was revised to generate a new strategy consisting of fewer analytical stages and resulting in a higher yield of N-terminal peptides. The improved method utilised the same N-blocking and proteolysis steps as in the initial method (steps 1 and 2), however, the biotinylation and subsequent streptavidin coupling steps were replaced with a single step, consisting of the removal of internal peptides via coupling to NHS-activated Sepharose. This enhanced protocol results in a substantially increased amount of material available for MS analysis and led to a ten-fold increase in the number of protein identifications when applied to mouse skeletal muscle. The two strategies for N-terminal purification are summarised in Figure 6.1.

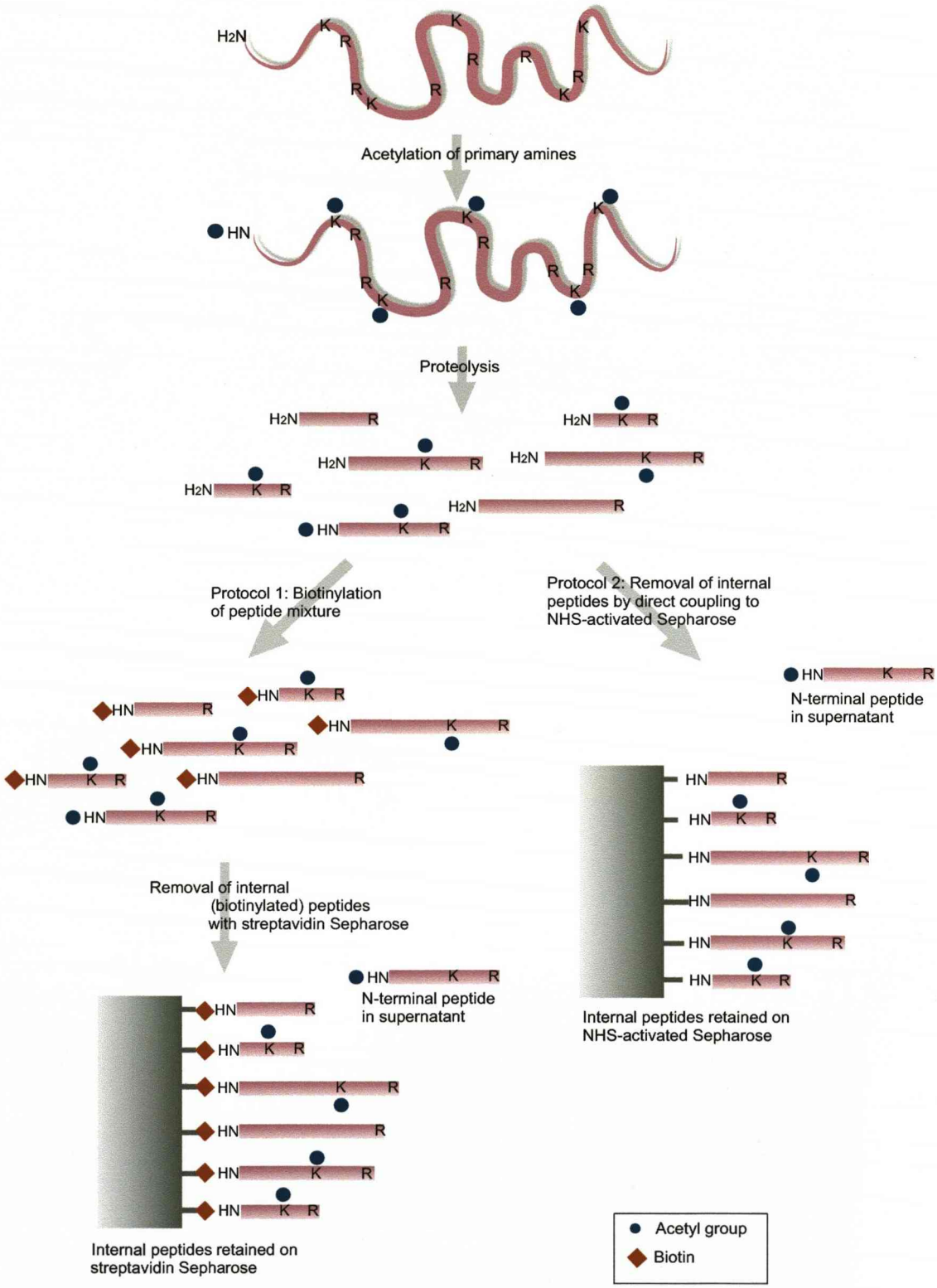


Figure 6.1. Comparison of the two protocols developed for N-terminal positional proteomics.
In both protocols, the intact protein mixture was acetylated in order to block exposed amino groups (α and ε). Proteolysis of the acetylated protein mixture generated blocked N-terminal and unblocked internal and C-terminal peptides. Following proteolysis, Protocol 1 (left) used NHS-Biotin to modify the exposed amino groups on the internal and C-terminal peptides. The biotinylated peptides were removed via incubation with streptavidin Sepharose. Alternatively, in Protocol 2 (right), the biotinylation stage was omitted and the acetylated peptide mixture was incubated directly with NHS-activated Sepharose without any further derivatisation steps.

When applied to the complex proteomes of mouse liver, *S. cerevisiae* and *E. coli*, the NHS-Sepharose strategy resulted in the identification of 333, 185 and 210 proteins respectively. Complete lists of N-terminal peptide identifications can be found in Supplementary data B. In addition to simplification, the positional proteomics strategy has the advantage of characterising the true N-terminal region of the identified protein, which has implications for genome annotation. Furthermore, by substituting the acetylation reagent used in the initial stage of the positional proteomics strategy to a labelled form of acetic anhydride ($C[^2H_3]CO)_2O$), it is possible to determine the N^{α}-acetylation status of the proteins identified. When applied to the three complex proteomes, this modified version of the N-terminal protocol confirmed the N^{α}-acetylation status of 98 mouse liver, 128 *S. cerevisiae* and 239 *E. coli* proteins. In accordance with published data 87% of the mouse proteins and 44% of the *S. cerevisiae* identified were N^{α}-acetylated *in vivo* (Polevoda and Sherman, 2000; Lee *et al.*, 1989b). In contrast to the eukaryotic samples, the majority of the *E. coli* proteins analysed were not N^{α}-acetylated *in vivo* (99%). The *E. coli* protein Elongation factor Tu is known to be N^{α}-acetylated *in vivo* (Arai *et al.*, 1980). However, this study also provided strong evidence (MS/MS sequence data) for the *in vivo* N^{α}-acetylation of copper resistance protein D, which is not reported in the literature. A complete representation of the data obtained on N^{α}-acetylation status can be found in Supplementary data B.

6.2 Limitations to the N-terminal 'positional proteomics' strategy

The limitations relating to the effectiveness of the approach are small. First, some tryptic fragments are too small or too large to be amenable to this analysis, although these can be accessed by alternative fragmentation methods such as the use of different proteases. Second, the presence of a small number of internal peptides in the preparation will result in the identification of some false positives. Internal peptides are present in the final preparation either as a result of incomplete coupling to NHS-activated Sepharose ('leakiness') or because they are modified in some way. The screen for HSA internal peptides in Section 4.5 of this thesis, showed that the presence of internals is extremely low compared with the true N-terminal peptide. Furthermore, it was not possible to confirm the validity of these peptides by MSMS data, making their presence in the sample uncertain. The occurrence of an N-blocked internal (Arg-C) peptide was highlighted in Section 3.9.10. This peptide was present in the *E. coli* N-terminal preparation and was derived from β-lactamase. This peptide contained a pyroglutamic acid at the N-terminal position which inhibited coupling to NHS-activated Sepharose. The observation that this peptide was in fact an Arg-C proteolysis product

suggests that this peptide was modified *in vitro* (after proteolysis) and does not indicate that the protein (β-lactamase) exists *in vivo* in a truncated form.

In terms of global proteomics, this strategy fails to deliver the amount of protein identifications as other "second generation" approaches such as MudPIT and COFRADIC. However, the utility of the methodology, to elucidate the true N-terminal region of a large number of protein species should not be underestimated. This ability alone lends itself to numerous other applications, including the characterisation of N-terminal PTM events and N-terminal susceptibility to many other enzymatic processes.

6.3 Application of positional proteomics to human plasma

MS-based proteomics is mostly used as a discovery tool in understanding disease. Substantial improvements in the sensitivity of analytical instrumentation and the availability of the human genome have made possible the discovery, identification and characterisation of proteins of diagnostic value. Recent reports have introduced different platforms for candidate based targeted proteomics which utilise stable isotope labelled, synthetic peptides as references for the identification and quantification of a selected group of target peptides. However, current proteomics techniques have failed to deliver robust strategies for clinical application, particularly for lower abundance biomarkers. Rather, the putative biomarkers identified through MS-based studies must be subjected to further validation using other high-throughput methods. This has led to a striking shortfall in the number of protein diagnostic tests emerging from proteomic studies. In order to bridge the gap between biomarker discovery and clinical use it will be necessary to develop novel approaches to target known biomarkers and quantitatively detect them in multiple clinical samples.

It was anticipated that N-terminal positional proteomics strategy would provide a method of reducing the complexity of the sample of choice for biomarker analysis, human plasma. However, application of the N-terminal strategy led to the identification of only 50 species, the majority of which were classical plasma proteins. The inability to reach deeper into the plasma proteome is a direct result of the large dynamic concentration range exhibited by this complex sample. Even with the utility of Protein Equalizer™ technology to normalise the protein concentrations within the plasma sample, it was not possible to identify more than 100 proteins by virtue of their N-terminal peptide. Although the method failed to deliver a comprehensive profile of the plasma proteome, the study provided some interesting insights into the nature of plasma protein N-termini. For instance, the interrogation of accurate mass

data from the Orbitrap instrument led to the identification of putative substrates for enzymes that remove dipeptides from the N-termini of proteins (dipeptidyl peptidases). Many of the classical plasma proteins showed evidence of N-terminal "trimming" including HSA, α-1-antichymotrypsin, apolipoprotein C and α-2-HS-glycoprotein. These proteins could serve as "catalytic reporters" in measuring the activity of dipeptidyl peptidases such as DPP IV and in turn provide insights into diseases associated with increased enzyme activity.

Although this study failed to identify a large number of previously identified candidate biomarkers, the ability to reproducibly identify the complement derived C3a anaphylatoxin is of substantial clinical relevance. This protein is up regulated in colorectal tumours (Habermann *et al.*, 2006) and the ability to identify this protein in multiple human plasma samples could provide the basis of a diagnostic screen for this putative biomarker.

There is evidence to suggest that the immune system may act as a biomarker for the presence, type and possibly even the stage of cancer (Finn, 2005). For this reason, it is possible that immunoglobulin profiling may provide insights into the molecular mechanisms of cancer progression (Tan, 2001). The ability of the N-terminal purification strategy to access the variable region of immunoglobulin molecules may provide a rapid method in which to screen these molecules for changes associated with disease.

6.4 MIDAR

In global strategies such as MudPIT, MS/MS analysis presents the rate limiting step. The amount of information obtained is limited due to constraints imposed by duty cycles and resolution trade offs. An ideal situation for global proteomics would be a single run of high-resolution, high mass accuracy MS data that could alone provide comprehensive proteome coverage without the requirement for MS/MS data. The AMT strategy discussed in Section 5.1.4 obviates the routine need for MS/MS and in turn reduces sample requirements whilst increasing throughput. A desirable attribute for an AMT based proteomics experiment is added parameters for database searching, which can be used along with the existing MS data to validate a match.

A second challenge to this study was to devise a new amino group labelling reagent to provide information regarding the chemical composition of the targeted species. The MIDAR strategy is a straight forward peptide labelling step that allows the quantification of lysine residues within a proteolytic peptide.

The MIDAR reagent was used to label a range of model peptides, purified proteins and biological samples. In all cases the labelling pattern gave unambiguous indication of the number of modified amino groups.

The logic behind MIDAR could also be applied using other reagents. For example, a simpler reagent, comprising a mixture of stable isotope labelled [$^{13}C_2$] and unlabelled acetic anhydride could also be used. In this instance, the "minus 1" signature would become "minus 2" and the analysis should otherwise be the same. However, this approach is less reliable, as unlabelled peptide might reflect the number of amino groups but also be compromised by incomplete labelling. Incomplete reaction would be less of an issue, if used in combination with a positional proteomics strategy as any incompletely acetylated peptide would be removed downstream. However, as a general strategy for determination of the frequency of lysine residues, using a reagent comprising two differentially labelled compounds that are both stable isotope labelled means that this complication is removed, as incomplete labelling should not influence the relative peak intensity of the "minus one" ion.

Partial metabolic labelling with a stable isotope labelled amino acid can also be used to determine the frequency of that amino acid in a peptide (Beynon and Pratt, 2005; Pratt *et. al.*, 2002). Information regarding the number of leucine residues in each peptide gave an enhancement of search efficiency of approximately one order of magnitude. The MIDAR approach is an experimental protocol conducted *in vitro* and has a more general applicability, whether used for a complete proteolytic digest or as part of a positional proteomics strategy. It is possible that other dual-labelled reagents could be developed to provide similar data on amino acid frequency. Coupled with accurate mass data and N-terminal positional proteomics, this additional parameter will enhance the identification process in proteomics experiments, and allow rapid proteome profiling that would be particularly valuable for large scale screening or comparative studies.

6.5 Informatics challenges

N-terminal regions of proteins are generally underrepresented in routine datasets generated by PMF (Meinnel and Giglione, 2008). It is highly likely that the lack of N-terminal peptide matches in published datasets is due to modification events that introduce heterogeneity at this region. N-terminal trimming, involving the removal of small peptides, generates new N-termini that are not representative of the sequences present in databases. Additionally, proteins that have undergone SP cleavage will be represented in the database in their

precursor form, making it difficult, if not impossible to identify the mature N-terminal by sequence similarity searching methods alone. A useful resource for positional proteomic strategies would be a range of databases containing sequentially truncated or 'trimmed' protein isoforms which would, in turn, allow for the identification of N-terminally trimmed proteins. This issue can partly be addressed using 'no enzyme' or 'nonspecific (half) cleavage' options that are available in current database search engines. However, searches done using these parameters result in significant increase in analysis time (Craig, 2003).

From a validation perspective, database search engines (for example, Mascot) will favour protein identifications that are supported by at least two peptide matches (Perkins *et al.*, 1999; Sun *et al.*, 2007). For this reason, "one-hit wonders", in which proteins are identified by only one MS/MS sequenced peptide, will not be classed as confident assignments in terms of their score. Statistical analyses can be performed on database search results to measure the likelihood of false-positive peptide identifications. Decoy databases consist of randomised or reversed sequences, which can be searched in order to estimate the number of false positives present in the results from the true databases. When searching MS/MS data sets generated from N-terminal preparations, each protein will ideally be represented by a single peptide. In this case, the entire database from a given species will act as a decoy database, as the majority of data represented will be from non N-terminal peptides. The added information gain obtained from the knowledge of the peptide location within the parent protein is sufficient, in most cases, to validate the assignment. Although, in the case of a poor quality MS/MS spectrum acquired from low abundance proteins, the confidence of identification will be poor and most likely insignificant. The fundamental problem is that currently MS search software does not incorporate tools to permit the inclusion of positional information. Subsequently, the added information gain relating to the position of the peptide is lost and will not directly contribute to the confidence of the match.

6.6 Future work

Selective or multiple reaction monitoring (MRM) in mass spectrometry is a method used to confirm, unambiguously, the presence of a compound in a complex sample. It is not only a highly specific method but also has very high sensitivity. The MRM approach has been applied to the quantitative analysis of tryptic peptides representing high and medium abundance proteins in human plasma (Anderson and Hunter, 2006).

Preliminary data reveals that targeting product ions originating from N-terminal peptides from an N-terminal preparation of human plasma, allows a large increase in sensitivity (work conducted by Gemma Davidson). It is anticipated that combining N-terminal enrichment of human plasma proteins with an MRM based approach will provide a powerful strategy for the detection of low abundance putative biomarkers in human plasma.

Although the enhanced (NHS-Sepharose) N-terminal method produced a higher yield of peptides than the original biotin/streptavidin method, there still remains concern regarding the losses of N-terminal yield throughout the post-proteolysis purification stages. It will be beneficial to investigate the effects of other immobilised reagents with a view to increasing the amount of material in the final preparation. Reagents such as CNBr and phenyl isothiocyanates are available as activated supports, which could be used to replace NHS-Sepharose in the final N-terminal purification step.

Absolute quantification of multiple proteins within a biological system is achievable through use of an artificial protein known as a QconCAT. The QconCAT strategy was developed in this laboratory (Beynon *et al.*, 2005) and provides a tailored approach for multiplexed protein quantification. Coupling of the N-terminal strategy to a QconCAT based approach will allow simultaneous simplification and quantification of a subset of proteins. A potential use for this combined technique is quantification of the immunoglobulin repertoire found within human individuals. To achieve this, a QconCAT would be designed using N-terminal peptide sequences (Arg-C), of a selected group of immunoglobulins, to quantify the changes between individuals. As previously described, immunoglobulin profiling may provide insights into cancer and autoimmune disorders. Therefore, by linking the QconCAT quantification strategy to a method that targets the specific region of interest (variable region at the N-termini of immunoglobulin molecules); this combined strategy could form the basis of a useful clinical application.

6.7 Concluding remarks

The work in this thesis demonstrates that the N-terminal positional proteomics strategy is an elegant technique for simplification and characterisation of complex proteomes. In addition to reducing sample complexity, N-terminal positional proteomics can also be used for the investigation of *in vitro* processing events such as SP removal, NME and N^{α}-acetylation. The challenge for this technique, and many other global proteomic strategies, is the characterisation of the least abundant portion of the proteome (<90%).

Combining targeted analysis of N-terminal protein regions with *in vitro* labelling using MIDAR to provide additional information (positional and compositional) for database searching is of substantial interest to the field of global proteomics. However, a fundamental problem associated with positional based peptide isolation strategies, is that currently MS search software does not incorporate tools to permit the inclusion of positional information. Consequently, this positional information is lost and will not directly contribute to the confidence of the match. In order to maximise the utility of positional proteomics approach, the next challenge is to develop enhanced applications for database searching allowing the inclusion of positional information.

7. References

7. REFERENCES

Adkins, J.N., Varnum, S.M., Auberry, K.J., *et al.* (2002) Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. *Mol Cell Proteomics* **1**, 947-955.

Aebersold, R., Anderson, L., Caprioli, R., Druker, B., Hartwell, L. and Smith, R. (2005) Perspective: a program to improve protein biomarker discovery for cancer. *J Proteome Res* **4**, 1104-1109.

Aggarwal, K., Choe, L.H. and Lee, K.H. (2005) Quantitative analysis of protein expression using amine-specific isobaric tags in *Escherichia coli* cells expressing rhsA elements. *Proteomics* **5**, 2297-2308.

Aggarwal, K., Choe, L.H. and Lee, K.H. (2006) Shotgun proteomics using the iTRAQ isobaric tags. *Brief Funct Genomic Proteomic* **5**, 112-120.

Ako, H., Foster, R.J. and Ryan, C.A. (1974) Mechanism of action of naturally occurring proteinase inhibitors. Studies with anhydrotrypsin and anhydrochymotrypsin purified by affinity chromatography. *Biochemistry* **13**, 132-139.

Altschul, S.F., Madden, T.L., Schaffer, A.A., *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.

Anderson, L. and Seilhamer, J. (1997) A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* **18**, 533-537.

Anderson, N.L. and Anderson, N.G. (2002) The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* **1**, 845-867.

Anderson, N.L., Polanski, M., Pieper, R., *et al.* (2004) The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol Cell Proteomics* **3**, 311-326.

Anderson, L. and Hunter, C.L. (2006) Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol Cell Proteomics* **5**, 573-588.

Arai, K., Clark, B.F., Duffy, L., *et al.* (1980) Primary structure of elongation factor Tu from *Escherichia coli*. *Proc Natl Acad Sci U S A* **77**, 1326-1330.

Arnold, R.J., Polevoda, B., Reilly, J.P. and Sherman, F. (1999) The action of N-terminal acetyltransferases on yeast ribosomal proteins. *J Biol Chem* **274**, 37035-37040.

Atkinson, A.J., Colburn, W.A., DeGruttola, V.G., *et al.* (2001) Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clin Pharmacol Ther* **69**, 89-95.

Baczynskyj, L., Bronson, G.E. and Kubiak, T.M. (1994) Application of thermally assisted electrospray ionization mass spectrometry for detection of noncovalent complexes of bovine serum albumin with growth hormone releasing factor and other biologically active peptides. *Rapid Commun Mass Spectrom* **8**, 280-286.

Baggerly, K.A., Morris, J.S. and Coombes, K.R. (2004) Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* **20**, 777-785.

Baldeschwieler, J.D. (1968) Ion cyclotron resonance spectroscopy. Cyclotron double resonance provides a new technique for the study of ion-molecule reaction mechanisms. *Science* **159**, 263-273.

Baldo, A., Sniderman, A.D., St-Luce, S., *et al.* (1993) The adipsin-acylation stimulating protein system and regulation of intracellular triglyceride synthesis. *J Clin Invest* **92**, 1543-1547.

Bateman, A., Solomon, S. and Bennett, H.P. (1990) Post-translational modification of bovine pro-opiomelanocortin. Tyrosine sulfation and pyroglutamate formation, a mass spectrometric study. *J Biol Chem* **265**, 22130-22136.

Beardsley, R.L. and Reilly, J.P. (2002) Optimization of guanidination procedures for MALDI mass mapping. *Anal Chem* **74**, 1884-1890.

Becker, J.M. and Wilchek, M. (1972) Inactivation by avidin of biotin-modified bacteriophage. *Biochim Biophys Acta* **264**, 165-170.

Becker, J.M., Wilchek, M. and Katchalski, E. (1971) Irreversible inhibition of biotin transport in yeast by biotinyl-p-nitrophenyl ester. *Proc Natl Acad Sci U S A* **68**, 2604-2607.

Belov, M.E., Anderson, G.A., Angell, N.H., *et al.* (2001) Dynamic range expansion applied to mass spectrometry based on data-dependent selective ion ejection in capillary liquid chromatography fourier transform ion cyclotron resonance for enhanced proteome characterization. *Anal Chem* **73**, 5052-5060.

Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783-795.

Betancourt, L., Besada, V., Gonzalez, L.J., *et al.* (2001) Selective isolation and identification of N-terminal blocked peptides from tryptic protein digests. *Journal of Peptide Research* **57**, 345-353.

Beynon, R.J., Doherty, M.K., Pratt, J.M. and Gaskell, S.J. (2005) Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides rates. *Nat Methods* **2**, 587-589

Beynon, R.J. (2005) A simple tool for drawing proteolytic peptide maps. *Bioinformatics* **21**, 674-675.

Beynon, R.J. and Pratt, J.M. (2005) Metabolic labeling of proteins for proteomics. *Mol Cell Proteomics* **4**, 857-872.

Bjellqvist, B., Ek, K., Righetti, P.G., *et al.* (1982) Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *J Biochem Biophys Methods* **6**, 317-339.

Blagoev, B., Kratchmarova, I., Ong, S.E., Nielsen, M., Foster, L.J. and Mann, M. (2003) A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nat Biotechnol* **21**, 315-318.

Blattner, F.R., Plunkett, G., Bloch, C.A., *et al.* (1997) The Complete Genome Sequence of *Escherichia coli* K-12. *Science* **277**, 1453-1462.

Bode, W., Schwager, P. and Huber, R. (1978) The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. The refined crystal structures of the bovine trypsinogen-pancreatic trypsin inhibitor complex and of its ternary complex with Ile-Val at 1.9 Å resolution. *J Mol Biol* **118**, 99-112.

Boutin, J.A. (1997) Myristoylation. *Cellular Signalling* **9**, 15-35.

Brancia, F.L., Butt, A., Beynon, R.J., Hubbard, S.J., Gaskell, S.J. and Oliver, S.G. (2001) A combination of chemical derivatisation and improved bioinformatic tools optimises protein identification for proteomics. *Electrophoresis* **22**, 552-559.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J. and Bork, P. (2002) Alternative splicing and genome complexity. *Nat Genet* **30**, 29-30.

Brown, J.R. and Hartley, B.S. (1966) Location of disulphide bridges by diagonal paper electrophoresis. The disulphide bridges of bovine chymotrypsinogen A. *Biochem J* **101**, 214-228.

Browne, C.A., Bennett, H.P. and Solomon, S. (1982) The isolation of peptides by high-performance liquid chromatography using predicted elution positions. *Anal Biochem* **124**, 201-208.

Cárdenas, M.S., Van der Heeft, E. and de Jong, A.P. (1997) On-line derivatization of peptides for improved sequence analysis by micro-column liquid chromatography coupled with electrospray ionization-tandem mass spectrometry. *Rapid Commun Mass Spectrom* **11**, 1271-1278.

Carter, W.A. (1981) Binding of human interferons to immobilized albumin. *Methods Enzymol* **78**, 576-582.

Chao, C.C., Ma, Y.S. and Stadtman, E.R. (1997) Modification of protein surface hydrophobicity and methionine oxidation by oxidative systems. *Proc Natl Acad Sci U S A* **94**, 2969-2974.

Charbaut, E., Redeker, V., Rossier, J. and Sobel, A. (2002) N-terminal acetylation of ectopic recombinant proteins in *Escherichia coli*. *FEBS Lett* **529**, 341-345.

Chen, G., Gharib, T.G., Huang, C.C., *et al.* (2002) Discordant protein and mRNA expression in lung adenocarcinomas. *Mol Cell Proteomics* **1**, 304-313.

Chen, Y., Sprung, R., Tang, Y., *et al.* (2007) Lysine propionylation and butyrylation are novel post-translational modifications in histones. *Mol Cell Proteomics* **6**, 812-819.

Chock, P.B., Rhee, S.G. and Stadtman, E.R. (1980) Interconvertible enzyme cascades in cellular regulation. *Annu Rev Biochem* **49**, 813-843.

Choe, L., D'Ascenzo, M., Relkin, N.R., *et al.* (2007) 8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease. *Proteomics* **7**, 3651-3660.

Choudhary, J.S., Blackstock, W.P., Creasy, D.M. and Cottrell, J.S. (2001) Matching peptide mass spectra to EST and genomic DNA databases. *Trends Biotechnol* **19**, S17-22.

Christianson, D.W. and Lipscomb, W.N. (1989) Carboxypeptidase-A. *Accounts Chem Res* **22**, 62-69.

Cianflone, K., Xia, Z. and Chen, L.Y. (2003) Critical review of acylation-stimulating protein physiology in humans and rodents. *Biochim Biophys Acta* **1609**, 127-143.

Cianflone, K.M., Sniderman, A.D., Walsh, M.J., Vu, H.T., Gagnon, J. and Rodriguez, M.A. (1989) Purification and characterization of acylation stimulating protein. *J Biol Chem* **264**, 426-430.

Ciechanover, A. and Ben-Saadon, R. (2004) N-terminal ubiquitination: more protein substrates join in. *Trends Cell Biol* **14**, 103-106.

Ciechanover, A. and Schwartz, A.L. (1989) How are substrates recognized by the ubiquitin-mediated proteolytic system? *Trends Biochem Sci* **14**, 483-488.

Cong, Y.S., Fan, E. and Wang, E. (2006) Simultaneous proteomic profiling of four different growth states of human fibroblasts, using amine-reactive isobaric tagging reagents and tandem mass spectrometry. *Mech Ageing Dev* **127**, 332-343.

Conrads, T.P., Anderson, G.A., Veenstra, T.D., Pasa-Tolic, L. and Smith, R.D. (2000) Utility of accurate mass tags for proteome-wide protein identification. *Anal Chem* **72**, 3349-3354.

Corthals, G.L., Wasinger, V.C., Hochstrasser, D.F. and Sanchez, J.C. (2000) The dynamic range of protein expression: a challenge for proteomic research. *Electrophoresis* **21**, 1104-1115.

Creasy, D.M. and Cottrell, J.S. (2004) Unimod: Protein modifications for mass spectrometry. *Proteomics* **4**, 1534-1536.

Craig, R. and Beavis, R.C. (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom* **17**, 2310-2316.

Davies, D.R., Padlan, E.A. and Sheriff, S. (1990) Antibody-antigen complexes. *Annu Rev Biochem* **59**, 439-473.

de Bruijn, M.H. and Fey, G.H. (1985) Human complement component C3: cDNA coding sequence and derived primary structure. *Proc Natl Acad Sci U S A* **82**, 708-712.

Derman, A.I. and Beckwith, J. (1991) *Escherichia coli* alkaline phosphatase fails to acquire disulfide bonds when retained in the cytoplasm. *J Bacteriol* **173**, 7719-7722.

Doherty, M.K., McLean, L., Hayter, J.R., *et al.* (2004) The proteome of chicken skeletal muscle: changes in soluble protein expression during growth in a layer strain. *Proteomics* **4**, 2082-2093.

Doolittle, R.F. (1984) Fibrinogen and fibrin. *Annu Rev Biochem* **53**, 195-229.

Dormeyer, W., Mohammed, S., Breukelen, B.V., Krijgsveld, J. and Heck, A.J.R. (2007) Targeted Analysis of Protein Termini. *J. Proteome Res.* **6**, 4634-4645.

Dugaiczyk, A., Law, S.W. and Dennison, O.E. (1982) Nucleotide sequence and the encoded amino acids of human serum albumin mRNA. *Proc Natl Acad Sci U S A* **79**, 71-75.

Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. and Trent, J.M. (1999) Expression profiling using cDNA microarrays. *Nat Genet* **21**, 10-14.

Durinx, C., Lambeir, A.M., Bosmans, E., *et al.* (2000) Molecular characterization of dipeptidyl peptidase activity in serum: soluble CD26/dipeptidyl peptidase IV is responsible for the release of X-Pro dipeptides. *Eur J Biochem* **267**, 5608-5613.

Edman, P. (1949) A method for the determination of amino acid sequence in peptides. *Arch Biochem* **22**, 475.

Edman, P. and Begg, G. (1967) A protein sequenator. *Eur J Biochem* **1**, 80-91.

Emanuelsson, O., Nielsen, H., Brunak, S. and von Heijne, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* **300**, 1005-1016.

Engen, J.R., Bradbury, E.M. and Chen, X. (2002) Using stable-isotope-labeled proteins for hydrogen exchange studies in complex mixtures. *Anal Chem* **74**, 1680-1686.

Espagne, C., Martinez, A., Valot, B., Meinnel, T. and Giglione, C. (2007) Alternative and effective proteomic analysis in Arabidopsis. *Proteomics* **7**, 3788-3799.

Everley, P.A., Krijgsveld, J., Zetter, B.R. and Gygi, S.P. (2004) Quantitative cancer proteomics: stable isotope labeling with amino acids in cell culture (SILAC) as a tool for prostate cancer research. *Mol Cell Proteomics* **3**, 729-735.

Ewing, B. and Green, P. (2000) Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet* **25**, 232-234.

Farries, T.C., Harris, A., Auffret, A.D. and Aitken, A. (1991) Removal of N-acetyl groups from blocked peptides with acylpeptide hydrolase. Stabilization of the enzyme and its application to protein sequencing. *Eur J Biochem* **196**, 679-685.

Ficarro, S.B., McCleland, M.L., Stukenberg, P.T., *et al.* (2002) Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nat Biotechnol* **20**, 301-305.

Finn, O.J. (2005) Immune response as a biomarker for cancer detection and a lot more. *N Engl J Med* **353**, 1288-1290.

Flinta, C., Persson, B., Jornvall, H. and von Heijne, G. (1986) Sequence determinants of cytosolic N-terminal protein processing. *Eur J Biochem* **154**, 193-196.

Frottin, F., Martinez, A., Peynot, P., *et al.* (2006) The proteomics of N-terminal methionine cleavage. *Mol Cell Proteomics* **5**, 2336-2349.

Gevaert, K., Goethals, M., Martens, L., *et al.* (2003) Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat Biotech* **21**, 566-569.

Gevaert, K., Van Damme, J., Goethals, M., *et al.* (2002) Chromatographic Isolation of Methionine-containing Peptides for Gel-free Proteome Analysis: Identification Of More Than 800 *Escherichia Coli* Proteins. *Mol Cell Proteomics* **1**, 896-903.

Gevaert, K., Ghesquière, B., Staes, A., *et al.* (2004) Reversible labeling of cysteine-containing peptides allows their specific chromatographic isolation for non-gel proteome studies. *Proteomics* **4**, 897-908.

Gevaert, K., Van Damme, P., Ghesquiere, B. and Vandekerckhove, J. (2006) Protein processing and other modifications analyzed by diagonal peptide chromatography. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1764**, 1801-1810.

Ghaemmaghami, S., Huh, W.-K., Bower, K., *et al.* (2003) Global analysis of protein expression in yeast. *Nature* **425**, 737-741.

Giglione, C., Pierre, M. and Meinnel, T. (2000) Peptide deformylase as a target for new generation, broad spectrum antimicrobial agents. *Mol Microbiol* **36**, 1197-1205.

Giglione, C., Vallon, O. and Meinnel, T. (2003) Control of protein life-span by N-terminal methionine excision. *Embo J* **22**, 13-23.

Goffeau, A., Barrell, B.G., Bussey, H., *et al.* (1996) Life with 6000 Genes. *Science* **274**, 546-567.

Guerrera, I.C. and Kleiner, O. (2005) Application of mass spectrometry in proteomics. *Biosci Rep* **25**, 71-93.

Granger, J., Siddiqui, J., Copeland, S. and Remick, D. (2005) Albumin depletion of human plasma also removes low abundance proteins including the cytokines. *Proteomics* **5**, 4713-4718.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H. and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotech* **17**, 994-999.

Gygi, S.P., Rochon, Y., Franza, B.R. and Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* **19**, 1720-1730.

Hagihara, M., Ohhashi, M. and Nagatsu, T. (1987) Activities of dipeptidyl peptidase II and dipeptidyl peptidase IV in mice with lupus erythematosus-like syndrome and in patients with lupus erythematosus and rheumatoid arthritis. *Clin Chem* **8**, 1463-1465.

Habermann, J.K., Roblick, U.J., Luke, B.T., *et al.* (2006) Increased Serum Levels of Complement C3a Anaphylatoxin Indicate the Presence of Colorectal Tumors. *Gastroenterology* **131**, 1020-1029.

Hardt, M., Witkowska, H.E., Webb, S., *et al.* (2005) Assessing the effects of diurnal variation on the composition of human parotid saliva: quantitative analysis of native peptides using iTRAQ reagents. *Anal Chem* **77**, 4947-4954.

Hayter, J.R., Robertson, D.H., Gaskell, S.J. and Beynon, R.J. (2003) Proteome analysis of intact proteins in complex mixtures. *Mol Cell Proteomics* **2**, 85-95.

Hellerstein, M.K. and Neese, R.A. (1992) Mass isotopomer distribution analysis: a technique for measuring biosynthesis and turnover of polymers. *Am J Physiol Endocrinol Metab* **263**, E988-1001.

Henzel, W.J., Watanabe, C. and Stults, J.T. (2003) Protein identification: the origins of peptide mass fingerprinting. *Journal of the American Society for Mass Spectrometry* **14**, 931-942.

Hiller, K., Grote, A., Scheer, M., Munch, R. and Jahn, D. (2004) PrediSi: prediction of signal peptides and their cleavage positions. *Nucl. Acids Res.* **32**, W375-379.

Hinerfeld, D., Innamorati, D., Pirro, J. and Tam, S.W. (2004) Serum/Plasma Depletion with Chicken Immunoglobulin Y Antibodies for Proteomic Analysis from Multiple Mammalian Species. *J Biomol Tech* **15**, 184-190.

Hino, M., Fuyamada, H., Hayakawa, T., Nagatsu, T. and Oya, H. (1976) X-Prolyl dipeptidyl-aminopeptidase activity, with X-proline p-nitroanilides as substrates, in normal and pathological human sera. *Clin Chem* **22**, 1256-1261.

Hipple, J.A., Sommer, H. and Thomas, H.A. (1950) The Omegatron. *Phys Rev* **78**, 332-332.

Hirel, P.H., Schmitter, M.J., Dessen, P., Fayat, G. and Blanquet, S. (1989) Extent of N-terminal methionine excision from *Escherichia coli* proteins is governed by the side-chain length of the penultimate amino acid. *Proc Natl Acad Sci U S A* **86**, 8247-8251.

Honjo, T. and Habu, S. (1985) Origin of immune diversity: genetic variation and selection. *Annu Rev Biochem* **54**, 803-830.

Hoheisel, J.D. (2006) Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet* **7**, 200-210.

Hopsu-Havu, V.K. and Glenner, G.G. (1966) A new dipeptide naphthylamidase hydrolyzing glycyl-prolyl-beta-naphthylamide. *Histochemie* **7**, 197-201.

Horning, E.C., Carroll, D.I., Dzidic, I., *et al.* (1977) Development and use of analytical systems based on mass spectrometry. *Clin Chem* **23**, 13-21.

Hu, L., Ye, M., Jiang, X., Feng, S. and Zou, H. (2007) Advances in hyphenated analytical techniques for shotgun proteome and peptidome analysis. *Analytica Chimica Acta* **598**, 193-204.

Hugli, T.E. (1975) Human anaphylatoxin (C3a) from the third component of complement. Primary structure. *J Biol Chem* **250**, 8293-8301.

Huh, W.-K., Falvo, J.V., Gerke, L.C., *et al.* (2003) Global analysis of protein localization in budding yeast. *Nature* **425**, 686-691.

Hunter, T.C., Yang, L., Zhu, H., Majidi, V., Bradbury, E.M. and Chen, X. (2001) Peptide mass mapping constrained with stable isotope-tagged peptides for identification of protein mixtures. *Anal Chem* **73**, 4891-4902.

International Human Genome Sequencing Consortium.(2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

International Human Genome Sequencing Consortium.(2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945.

Issaq, H.J., Xiao, Z. and Veenstra, T.D. (2007) Serum and Plasma Proteomics. *Chem. Rev* **8**, 3601-3620..

Jackson, R.C. and Blobel, G. (1977) Post-translational cleavage of presecretory proteins with an extract of rough microsomes from dog pancreas containing signal peptidase activity. *Proc Natl Acad Sci U S A* **74**, 5598-5602.

Janssen, B.J., Huizinga, E.G., Raaijmakers, H.C., *et al.* (2005) Structures of complement component C3 provide insights into the function and evolution of immunity. *Nature* **437**, 505-511.

Johnson, J.M., Castle, J., Garrett-Engele, P., *et al.* (2003) Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays. *Science* **302**, 2141-2144.

Jonscher, K., Currie, G., McCormack, A.L. and Yates, J.R., 3rd (1993) Matrix-assisted laser desorption of peptides and proteins on a quadrupole ion trap mass spectrometer. *Rapid Commun Mass Spectrom* **7**, 20-26.

Kaiser, R.E., Jr., Williams, J.D., Lammert, S.A., Cooks, R.G. and Zakett, D. (1991) Thermospray liquid chromatography-mass spectrometry with a quadrupole ion trap mass spectrometer. *J Chromatogr* **562**, 3-11.

Karas, M. and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* **60**, 2299-2301.

Kasai, K. (1992) Trypsin and affinity chromatography. *J Chromatogr* **597**, 3-18.

Kenny, A.J., Booth, A.G., George, S.G., *et al.* (1976) Dipeptidyl peptidase IV, a kidney brush-border serine peptidase. *Biochem J* **157**, 169-182.

Kenyon, G.L., DeMarini, D.M., Fuchs, E., *et al.* (2002) Defining the mandate of proteomics in the post-genomics era: workshop report. *Mol Cell Proteomics* **1**, 763-780.

Klee, E.W. and Ellis, L.B. (2005) Evaluating eukaryotic secreted protein prediction. *BMC Bioinformatics* **6**, 256.

Klemenc, S. (2002) 4-Dimethylaminopyridine as a catalyst in heroin synthesis. *Forensic Sci Int* **129**, 194-199.

Krebs, E.G. and Beavo, J.A. (1979) Phosphorylation-Dephosphorylation of Enzymes. *Annual Review of Biochemistry* **48**, 923-959.

Krebs, H.A. (1950) Chemical composition of blood plasma and serum. *Annu Rev Biochem* **19**, 409-430.

Krishna, R.G., Chin, C.C. and Wold, F. (1991) N-terminal sequence analysis of N alpha-acetylated proteins after unblocking with N-acylaminoacyl-peptide hydrolase. *Anal Biochem* **199**, 45-50.

Krishna, R.G. and Wold, F. (1993) Post-translational modification of proteins. *Adv Enzymol Relat Areas Mol Biol* **67**, 265-298.

Krokhin, O.V., Craig, R., Spicer, V., *et al.* (2004) An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: its application to protein peptide mapping by off-line HPLC-MALDI MS. *Mol Cell Proteomics* **3**, 908-919.

Kuhn, K., Prinz, T., Schafer, J., *et al.* (2005) Protein sequence tags: a novel solution for comparative proteomics. *Proteomics* **5**, 2364-2368.

Kuhn, K., Thompson, A., Prinz, T., *et al.* (2003) Isolation of N-terminal protein sequence tags from cyanogen bromide cleaved proteins as a novel approach to investigate hydrophobic proteins. *J Proteome Res* **2**, 598-609.

Kumar, A., Agarwal, S., Heyman, J.A., *et al.* (2002) Subcellular localization of the yeast proteome. *Genes Dev* **16**, 707-719.

Laemmli, U.K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680-685.

Lander, E.S., Linton, L.M., Birren, B., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

Langen, H., Takacs, B., Evers, S., *et al.* (2000) Two-dimensional map of the proteome of Haemophilus influenzae. *Electrophoresis* **21**, 411-429.

Lathe, G.H. and Ruthven, C.R. (1956) The separation of substances and estimation of their relative molecular sizes by the use of columns of starch in water. *Biochem J* **62**, 665-674.

Laursen, R.A. (1971) Solid-phase Edman degradation. An automatic peptide sequencer. *Eur J Biochem* **20**, 89-102.

Larkin, M.A., Blackshields, G., Brown, N.P., *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **21**, 2947-2948.

Lee, F.J., Lin, L.W. and Smith, J.A. (1989a) N^{α}-acetyltransferase deficiency alters protein synthesis in *Saccharomyces cerevisiae*. *FEBS Letters* **256**, 139-142.

Lee, F.J., Lin, L.W. and Smith, J.A. (1989b) N^{α}-acetylation is required for normal growth and mating of *Saccharomyces cerevisiae*. *J Bacteriol* **171**, 5795-5802.

Link, A.J., Eng, J., Schieltz, D.M., *et al.* (1999) Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* **17**, 676-682.

Link, A.J., Robison, K. and Church, G.M. (1997) Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12. *Electrophoresis* **18**, 1259-1313.

Lipton, M.S., Romine, M.F., Monroe, M.E., *et al.* (2006) AMT tag approach to proteomic characterization of *Deinococcus radiodurans and Shewanella oneidensis*. *Methods Biochem Anal* **49**, 113-134.

Lipton, M.S., Romine, M.F., Monroe, M.E., *et al.*, (2006) AMT tag approach to proteomic characterization of *Deinococcus radiodurans* and *Shewanella oneidensis*. *Methods Biochem Anal* **49**, 113-134.

Litman, G.W., Rast, J.P., Shamblott, M.J., *et al.* (1993) Phylogenetic diversification of immunoglobulin genes and the antibody repertoire. *Mol Biol Evol* **10**, 60-72.

Liu, P., Feasley, C.L. and Regnier, F.E. (2004) Optimization of diagonal chromatography for recognizing post-translational modifications. *J Chromatogr A* **1047**, 221-228.

Liu, T., Belov, M.E., Jaitly, N., Qian, W.J. and Smith, R.D. (2007) Accurate mass measurements in proteomics. *Chem Rev* **107**, 3621-3653.

Lojda, Z. (1979) Studies on dipeptidyl(amino)peptidase IV (glycyl-proline naphthylamidase). II. Blood vessels. *Histochemistry* **59**, 153-166.

Magnuson, B.A., Raju, R.V.S., Moyana, T.N. and Sharma, R.K. (1995) Increased N-Myristoyltransferase Activity Observed in Rat and Human Colonic Tumors. *J. Natl. Cancer Inst.* **87**, 1630-1635.

Mant, C.T. and Hodges, R.S. (2002) Reversed-phase liquid chromatography as a tool in the determination of the hydrophilicity/hydrophobicity of amino acid side-chains at a ligand-receptor interface in the presence of different aqueous environments: II. Effect of varying peptide ligand hydrophobicity. *J Chromatogr A* **972**, 61-75.

Marekov, L.N. and Steinert, P.M. (2003) Charge derivatization by 4-sulfophenyl isothiocyanate enhances peptide sequencing by post-source decay matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Journal of Mass Spectrometry* **38**, 373-377.

Maroux, S., Baratti, J. and Desnuelle, P. (1971) Purification and Specificity of Porcine Enterokinase. *J. Biol. Chem.* **246**, 5031-5039.

Marshall, A.G., Hendrickson, C.L. and Jackson, G.S. (1998) Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom Rev* **17**, 1-35.

Martinez, A., Traverso, J.A., Valot, B., *et al.* (2008) Extent of N-terminal modifications in cytosolic proteins from eukaryotes. *Proteomics* **8**, 2809-2831.

Martinovic, S., Veenstra, T.D., Anderson, G.A., Pasa-Tolic, L. and Smith, R.D. (2002) Selective incorporation of isotopically labeled amino acids for identification of intact proteins on a proteome-wide level. *J Mass Spectrom* **37**, 99-107.

Matthews, L.R., Vaglio, P., Reboul, J., *et al.* (2001) Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved Protein-Protein Interactions or "Interologs". *Genome Res.* **11**, 2120-2126.

McComb, J.M., McMaster, E.A., MacKenzie, G. and Adgey, A.A. (1984) Myoglobin and creatine kinase in acute myocardial infarction. *Br Heart J* **51**, 189-194.

McDonald, L. and Beynon, R.J. (2006) Positional proteomics: preparation of amino-terminal peptides as a strategy for proteome simplification and characterization. *Nat Protoc* **1**, 1790-1798.

McDonald, L., Robertson, D.H., Hurst, J.L. and Beynon, R.J. (2005) Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides. *Nat Methods* **2**, 955-957.

Meinnel, T. and Giglione, C. (2008) Tools for analyzing and predicting N-terminal protein modifications. *Proteomics* **8**, 626-649.

Mikami, T. and Takao, T. (2007) Selective Isolation of N-Blocked Peptides by Isocyanate-Coupled Resin. *Anal. Chem.* **79**, 7910-7915.

Miller, B.T. (1996) Acylation of Peptide Hydroxyl Groups with the Bolton-Hunter Reagent. *Biochem Biophys Res Commun* **218**, 377-382.

Mirzaei, H. and Regnier, F. (2005) Structure specific chromatographic selection in targeted proteomics. *J Chromatogr B Analyt Technol Biomed Life Sci* **817**, 23-34.

Moerschell, R.P., Hosokawa, Y., Tsunasawa, S. and Sherman, F. (1990) The specificities of yeast methionine aminopeptidase and acetylation of amino-terminal methionine *in vivo*. Processing of altered iso-1-cytochromes c created by oligonucleotide transformation. *J Biol Chem* **265**, 19638-19643.

Moss, G.P., Smith, P.A.S. and Tavernier, D. (1995) Glossary of Class Names of Organic-Compounds and Reactive Intermediates Based on Structure. *Pure and Applied Chemistry* **67**, 1307-1375.

Nakazawa, T., Yamaguchi, M., Nishida, K., *et al.* (2004) Enhanced responses in matrix-assisted laser desorption/ionization mass spectrometry of peptides derivatized with arginine via a C-terminal oxazolone. *Rapid Commun Mass Spectrom* **18**, 799-807.

Nielsen, H., Engelbrecht, J., von Heijne, G. and Brunak, S. (1996) Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. *Proteins* **24**, 165-177.

O'Farrell, P.H. (1975) High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **250**, 4007-4021.

Omenn, G.S., States, D.J, Adamski, M., *et al.* (2005) Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **5**, 3226-3245.

Ong, S.E., Blagoev, B., Kratchmarova, I., *et al.* (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1**, 376-386.

Opiteck, G.J., Lewis, K.C., Jorgenson, J.W. and Anderegg, R.J. (1997) Comprehensive on-line LC/LC/MS of proteins. *Anal Chem* **69**, 1518-1524.

Ostrove, S. (1990) Affinity chromatography: general methods. *Methods Enzymol* **182**, 357-371.

Overbeek, R., Larsen, N., Walunas, T., *et al.* (2003) The ERGO™ genome analysis and discovery system. *Nucl. Acids Res.* **31**, 164-171.

Palagi, P.M., Hernandez, P., Walther, D. and Appel, R.D. (2006) Proteome informatics I: bioinformatics tools for processing experimental data. *Proteomics* **6**, 5435-5444.

Pan, S., Gu, S., Bradbury, E.M. and Chen, X. (2003) Single peptide-based protein identification in human proteome through MALDI-TOF MS coupled with amino acids coded mass tagging. *Anal Chem* **75**, 1316-1324.

Pappin, D.J. (2003) Peptide mass fingerprinting using MALDI-TOF mass spectrometry. *Methods Mol Biol* **211**, 211-219.

Pappin, D.J., Hojrup, P. and Bleasby, A.J.(1993) Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol* **3**, 327-32.

Perkins, D.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551-3567.

Perryman, M.B., Knell, J.D. and Roberts, R. (1984) Carboxypeptidase-catalyzed hydrolysis of C-terminal lysine: mechanism for *in vivo* production of multiple forms of creatine kinase in plasma. *Clin Chem* **30**, 662-664.

Peters, T. (1995) All About Albumin: Biochemistry, Genetics, and Medical Applications Academic Press Ltd.

Picotti, P., Aebersold, R. and Domon, B. (2007) The Implications of Proteolytic Background for Shotgun Proteomics. *Mol Cell Proteomics* **6**, 1589-1598.

Pieper, R., Su, Q., Gatlin, C.L., *et al.* (2003) Multi-component immunoaffinity subtraction chromatography: An innovative step towards a comprehensive survey of the human plasma proteome. *Proteomics* **3**, 422-432.

Polanski, M. and Anderson, N.L. (2006) A list of candidate cancer biomarkers for targeted proteomics. *Biomarker Insights* **1**, 1-48.

Polevoda, B., Norbeck, J., Takakura, H., Blomberg, A. and Sherman, F. (1999) Identification and specificities of N-terminal acetyltransferases from *Saccharomyces cerevisiae*. *Embo J* **18**, 6155-6168.

Polevoda, B. and Sherman, F. (2000) N^{α} -terminal acetylation of eukaryotic proteins. *J Biol Chem* **275**, 36479-36482.

Polevoda, B. and Sherman, F. (2003) N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins. *J Mol Biol* **325**, 595-622.

Porath, J. and Flodin, P. (1959) Gel filtration: a method for desalting and group separation. *Nature* **183**, 1657-1659.

Pratt, J.M., Robertson, D.H., Gaskell, S.J., *et al.* (2002) Stable isotope labelling *in vivo* as an aid to protein identification in peptide mass fingerprinting. *Proteomics* **2**, 157-163.

Prchal, J.T., Cashman, D.P. and Kan, Y.W. (1986) Hemoglobin Long Island is Caused by a Single Mutation (Adenine to Cytosine) Resulting in a Failure to Cleave Amino-Terminal Methionine. *Proc Natl Acad Sci U S A* **83**, 24-27.

Quackenbush, J. (2006) Microarray analysis and tumor classification. *N Engl J Med* **354**, 2463-2472.

Rademacher, T.W., Parekh, R.B. and Dwek, R.A. (1988) Glycobiology. *Annu Rev Biochem* **57**, 785-838.

Raggiaschi, R., Gotta, S. and Terstappen, G.C. (2005) Phosphoproteome analysis. *Biosci Rep* **25**, 33-44.

Raida, M., Schulz-Knappe, P., Heine, G. and Forssmann, W.G. (1999) Liquid chromatography and electrospray mass spectrometric mapping of peptides from human plasma filtrate. *Journal of the American Society for Mass Spectrometry* **10**, 45-54.

Rajala, R.V., Radhi, J.M., Kakkar, R., Datla, R.S. and Sharma, R.K. (2000) Increased expression of N-myristoyltransferase in gallbladder carcinomas. *Cancer* **88**, 1992-1999.

Ransohoff, D.F. (2005) Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer* **5**, 142-149.

Redlitz, A., Tan, A.K., Eaton, D.L. and Plow, E.F. (1995) Plasma Carboxypeptidases as Regulators of the Plasminogen System. *Journal of Clinical Investigation* **96**, 2534-2538.

Reid, K.B. and Porter, R.R. (1981) The proteolytic activation systems of complement. *Annu Rev Biochem* **50**, 433-464.

Regnier, F.E. and Gooding, K.M. (1980) High-performance liquid chromatography of proteins. *Anal Biochem* **103**, 1-25.

Reynolds, K.J., Yao, X. and Fenselau, C. (2002) Proteolytic ^{18}O labeling for comparative proteomics: evaluation of endoprotease Glu-C as the catalytic agent. *J Proteome Res* **1**, 27-33.

Righetti, P.G. and Caravaggio, T. (1976) Isoelectric points and molecular weights of proteins. *J Chromatogr* **127**, 1-28.

Righetti, P.G., Castagna, A. and Herbert, B. (2001) Prefractionation techniques in proteome analysis. *Anal Chem* **11**, 320A-326A.

Righetti, P.G., Boschetti, E., Lomas, E. and Citterio, A. (2006) Protein Equalizer Technology : the quest for a "democratic proteome". *Proteomics* **6**, 3980-3992.

Righetti, P.G. and Boschetti, E. (2007) Sherlock Holmes and the proteome - a detective story. *FEBS Journal* **274**, 897-905.

Rivers, J., McDonald, L., Edwards, I.J. and Beynon, R.J. (2008) Asparagine deamidation and the role of higher order protein structure. *J Proteome Res* **7**, 921-927.

Robinson, A.B., Scotchler, J.W. and McKerrow, J.H. (1973) Rates of nonenzymatic deamidation of glutaminyl and asparaginyl residues in pentapeptides. *J Am Chem Soc* **95**, 8156-8159.

Roise, D., Theiler, F., Horvath, S.J., *et al.* (1988) Amphiphilicity is essential for mitochondrial presequence function. *Embo J* **7**, 649-653.

Romine, M.F., Elias, D.A., Monroe, M.E., *et al.*, (2004) Validation of Shewanella oneidensis MR-1 small proteins by AMT tag-based proteome analysis. *Omics* **8**, 239-254.

Ross, P.L., Huang, Y.N., Marchese, J.N., *et al.* (2004) Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **3**, 1154-1169.

Rout, M.P. and Field, M.C. (2001) Isolation and characterization of subnuclear compartments from *Trypanosoma brucei*. Identification of a major repetitive nuclear lamina component. *J Biol Chem* **276**, 38261-38271.

Ryskjaer, J., Deacon, C.F., Carr, R.D., *et al.* (2006) Plasma dipeptidyl peptidase-IV activity in patients with type-2 diabetes mellitus correlates positively with HbAlc levels, but is not acutely affected by food intake. *Eur J Endocrinol* **155**, 485-493.

Sabatini, D.D., Blobel, G., Nonomura, Y. and Adelman, M.R. (1971) Ribosome-membrane interaction: Structural aspects and functional implications. *Adv Cytopharmacol* **1**, 119-129.

Sadygov, R.G., Cociorva, D. and Yates, J.R., 3rd (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods* **1**, 195-202.

Schatz, P.J. and Beckwith, J. (1990) Genetic analysis of protein export in *Escherichia coli*. *Annu Rev Genet* **24**, 215-248.

Schmidt, B.Z. and Colten, H.R. (2000) Complement: a critical test of its biological importance. *Immunol Rev* **178**, 166-176.

Schuerenberg, M., Luebbert, C., Eickhoff, H., Kalkum, M., Lehrach, H. and Nordhoff, E. (2000) Prestructured MALDI-MS sample supports. *Anal Chem* **72**, 3436-3442.

Scigelova, M. and Makarov, A. (2006) Orbitrap mass analyzer--overview and applications in proteomics. *Proteomics* **6 Suppl 2**, 16-21.

Sechi, S. and Chait, B.T. (2000) A Method To Define the Carboxyl Terminal of Proteins. *Anal. Chem.* **72**, 3374-3378.

Senko, M. W., Beu, S. C. and McLafferty, F. W. (1995) Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions. *Journal of the American Society for Mass Spectrometry* **6**, 229-233.

Sennels, L., Salek, M., Lomas, L., Boschetti, E., Righetti, P.G. and Rappsilber, J. (2007) Proteomic Analysis of Human Blood Serum Using Peptide Library Beads. *J. Proteome Res.* **6**, 4055-4062.

Shapiro, A.L., Vinuela, E. and Maizel, J.V., Jr. (1967) Molecular weight estimation of polypeptide chains by electrophoresis in SDS-polyacrylamide gels. *Biochem Biophys Res Commun* **28**, 815-820.

Sharov, A.A., Dudekula, D.B. and Ko, M.S.H. (2005) Genome-wide assembly and analysis of alternative transcripts in mouse. *Genome Res.* **15**, 748-754.

Sharp, P.A. (1985) On the origin of RNA splicing and introns. *Cell* **42**, 397-400.

Shen, Y., Kim, J., Strittmatter, E.F., *et al.* (2005) Characterization of the human blood plasma proteome. *Proteomics* **5**, 4034-4045.

Shiio, Y. and Aebersold, R. (2006) Quantitative proteome analysis using isotope-coded affinity tags and mass spectrometry. *Nat. Protocols* **1**, 139-145.

Shively, J.E. (2000) The chemistry of protein sequence analysis. *Exs* **88**, 99-117.

Smith, V.F., Schwartz, B.L., Randall, L.L. and Smith, R.D. (1996) Electrospray mass spectrometric investigation of the chaperone SecB. *Protein Sci* **5**, 488-494.

Smith, R.D., Anderson, G.A., Lipton, M.S., *et al.,* (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* **2**, 513-523.

Smyth, M.J., Godfrey, D.I. and Trapani, J.A. (2001) A fresh look at tumor immunosurveillance and immunotherapy. *Nat Immunol* **2**, 293-299.

Snijders, A.P., de Vos, M.G. and Wright, P.C. (2005) Novel approach for peptide quantitation and sequencing based on 15N and 13C metabolic labeling. *J Proteome Res* **4**, 578-585.

Statland, B.E., Winkel, P. and Bokelund, H. (1973) Factors contributing to intra-individual variation of serum constituents. 2. Effects of exercise and diet on variation of serum constituents in healthy subjects. *Clin Chem* **19**, 1380-1383.

Stephens, W.E. (2001) A pulsed mass spectrometer with time dispersion. *Phys Rev* **69**, 691-702.

Stroud, R.M., Kossiakoff, A.A. and Chambers, J.L. (1977) Mechanisms of zymogen activation. *Annu Rev Biophys Bioeng* **6**, 177-193.

Sun, S., Meyer-Arendt, K., Eichelberger, B., *et al.* (2007) Improved Validation of peptide MS/MS assignments using spectral intensity prediction. *Mol Cell Proteomics* **6**, 1-17.

Tam, S.W., Pirro, J. and Hinerfeld, D. (2004) Depletion and fractionation technologies in plasma proteomic analysis. *Expert Review of Proteomics* **1**, 411-420.

Tan, E.M. (2001) Autoantibodies as reporters identifying aberrant cellular mechanisms in tumorigenesis. *J Clin Invest* **108**, 1411-1415.

Tanaka, S., Matsushita, Y., Yoshikawa, A. and Isono, K. (1989) Cloning and molecular characterization of the gene rimL which encodes an enzyme acetylating ribosomal protein L12 of *Escherichia coli* K12. *Mol Gen Genet* **217**, 289-293.

Tanaka, Y., Akiyama, H., Kuroda, T., *et. al.* (2006) A novel approach and protocol for discovering extremely low-abundance proteins in serum. *Proteomics* **6**, 4845-4855.

Taylor, A. (1993) Aminopeptidases: structure and function. *FABSEB J* **2**, 290-298.

Thevis, M., Ogorzalek Loo, R.R. and Loo, J.A. (2003) In-gel derivatization of proteins for cysteine-specific cleavages and their analysis by mass spectrometry. *J Proteome Res* **2**, 163-172.

Tirumalai, R.S., Chan, K.C., Prieto, D.A., Issaq, H.J., Conrads, T.P. and Veenstra, T.D. (2003) Characterization of the Low Molecular Weight Human Serum Proteome. *Mol Cell Proteomics* **2**, 1096-1103.

Utsumi, T., Sato, M., Nakano, K., Takemura, D., Iwata, H. and Ishisaka, R. (2001) Amino Acid Residue Penultimate to the Amino-terminal Gly Residue Strongly Affects Two Cotranslational Protein Modifications, N-Myristoylation and N-Acetylation. *J. Biol. Chem.* **276**, 10505-10513.

Van Sommeren, A.P.G., Machielsen, P.A.G.M. and Gribnau, T.C.J. (1993) Comparison of three activated agaroses for use in affinity chromatography: Effects on coupling performance and ligand leakage. *J Chromatogr A* **639**, 23-31.\

Van Wynsberghe, D., Noback, C. R., Carola, R. (1995) Human Anatomy and Physiology, Third Edition edn, McGraw-Hill.

Varshavsky, A. (1992) The N-end rule. *Cell* **69**, 725-735.

Varshavsky, A. (1997) The N-end rule pathway of protein degradation. *Genes Cells* **2**, 13-28.

Veenstra, T.D., Conrads, T.P. and Issaq, H.J. (2004) What to do with "one-hit wonders"? *Electrophoresis* **25**, 1278-1279.

Veenstra, T.D., Conrads, T.P., Hood, B.L., Avellino, A.M., Ellenbogen, R.G. and Morrison, R.S. (2005) Biomarkers: mining the biofluid proteome. *Mol Cell Proteomics* **4**, 409-418.

Villanueva, J., Martorella, A.J., Lawlor, K., *et al.* (2006) Serum peptidome patterns that distinguish metastatic thyroid carcinoma from cancer-free controls are unbiased by gender and age. *Mol Cell Proteomics* **5**, 1840-1852.

von Heijne, G., Steppuhn, J. and Herrmann, R.G. (1989) Domain structure of mitochondrial and chloroplast targeting peptides. *Eur J Biochem* **180**, 535-545.

Walhout, A.J.M. and Vidal, M. (2001) Protein interaction maps for model organisms. *Nat Rev Mol Cell Biol* **2**, 55-63.

Washburn, M.P., Wolters, D. and Yates, J.R., 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19**, 242-247.

Warwood, S., Mohammed, S., Cristea, I.M., *et. al.* (2006) Guanidination chemistry for qualitative and quantitative proteomics. *Rapid Commun Mass Spectrom* **20**, 3245-3256.

Wiese, S., Reidegeld, K.A., Meyer, H.E. and Warscheid, B. (2007) Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *Proteomics* **7**, 340-350.

Whitehouse, C.M., Dreyer, R.N., Yamashita, M. and Fenn, J.B. (1985) Electrospray interface for liquid chromatographs and mass spectrometers. *Anal. Chem.* **57**, 675-679.

Wilchek, M. and Bayer, E.A. (1988) The avidin-biotin complex in bioanalytical applications. *Anal Biochem* **171**, 1-32.

Wilchek, M. and Bayer, E.A. (1989) Avidin-biotin technology ten years on: has it lived up to its expectations? *Trends Biochem Sci* **14**, 408-412.

Wilkins, M.R., Pasquali, C., Appel, R.D., *et al.* (1996) From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology (N Y)* **14**, 61-65.

Wilm, M., Shevchenko, A., Houthaeve, T., *et al.* (1996) Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379**, 466-469.

Wilmarth, P.A., Tanner, S., Dasari, S., *et al.* (2006) Age-related changes in human crystallins determined from comparative analysis of post-translational modifications in young and aged lens: does deamidation contribute to crystallin insolubility? *J Proteome Res* **5**, 2554-2566.

Wong, S.S. (1991) Chemistry of protein conjugation and cross-linking, CRC Press Ltd.

Wu, W.W., Wang, G., Baek, S.J. and Shen, R.F. (2006) Comparative Study of Three Proteomic Quantitative Methods, DIGE, cICAT, and iTRAQ, Using 2D Gel- or LC-MALDI TOF/TOF. *J. Proteome Res.* **5**, 651-658.

Xiong, L., Andrews, D. and Regnier, F. (2003) Comparative proteomics of glycoproteins based on lectin selection and isotope coding. *J Proteome Res* **2**, 618-625.

Yamaguchi, M., Nakazawa, T., Kuyama, H., *et al.* (2005) High-throughput method for N-terminal sequencing of proteins by MALDI mass spectrometry. *Anal Chem* **77**, 645-651.

Yamaguchi, M., Oka, M., Nishida, K., *et al.* (2006) Enhancement of MALDI-MS spectra of C-terminal peptides by the modification of proteins via an active ester generated *in situ* from an oxazolone. *Anal Chem* **78**, 7861-7869.

Yamaguchi, M., Obama, T., Kuyama, H., *et al.* (2007) Specific isolation of N-terminal fragments from proteins and their high-fidelity *de novo* sequencing. *Rapid Commun Mass Spectrom* **21**, 3329-3336.

Yao, X., Freas, A., Ramirez, J., Demirev, P.A. and Fenselau, C. (2001) Proteolytic ^{18}O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal Chem* **73**, 2836-2842.

Yoshikawa, A., Isono, S., Sheback, A. and Isono, K. (1987) Cloning and nucleotide sequencing of the genes rimI and rimJ which encode enzymes acetylating ribosomal proteins S18 and S5 of *Escherichia coli* K12. *Mol Gen Genet* **209**, 481-488.

Young, N., Chang, Z. and Wishart, D.S. (2004) GelScape: a web-based server for interactively annotating, manipulating, comparing and archiving 1D and 2D gel images. *Bioinformatics* **20**, 976-978.

Zhang, J., Goodlett, D.R., Peskind, E.R., *et al.* (2005a) Quantitative proteomic analysis of age-related changes in human cerebrospinal fluid. *Neurobiol Aging* **26**, 207-227.

Zhang, Z., Edwards, P.J., Roeske, R.W. and Guo, L. (2005b) Synthesis and Self-Alkylation of Isotope-Coded Affinity Tag Reagents. *Bioconjugate Chem.* **16**, 458-464.

Zhu, H., Hunter, T.C., Pan, S., Yau, P.M., Bradbury, E.M. and Chen, X. (2002) Residue-specific mass signatures for the efficient detection of protein modifications by mass spectrometry. *Anal Chem* **74**, 1687-1694.

Zieske, L.R. (2006) A perspective on the use of iTRAQ reagent technology for protein complex and profiling studies. *J Exp Bot* **57**, 1501-1508.

Supplementary Data

A. Identification of mouse muscle proteins by PMF

B. N-terminal identifications

C. Non N-terminal identifications

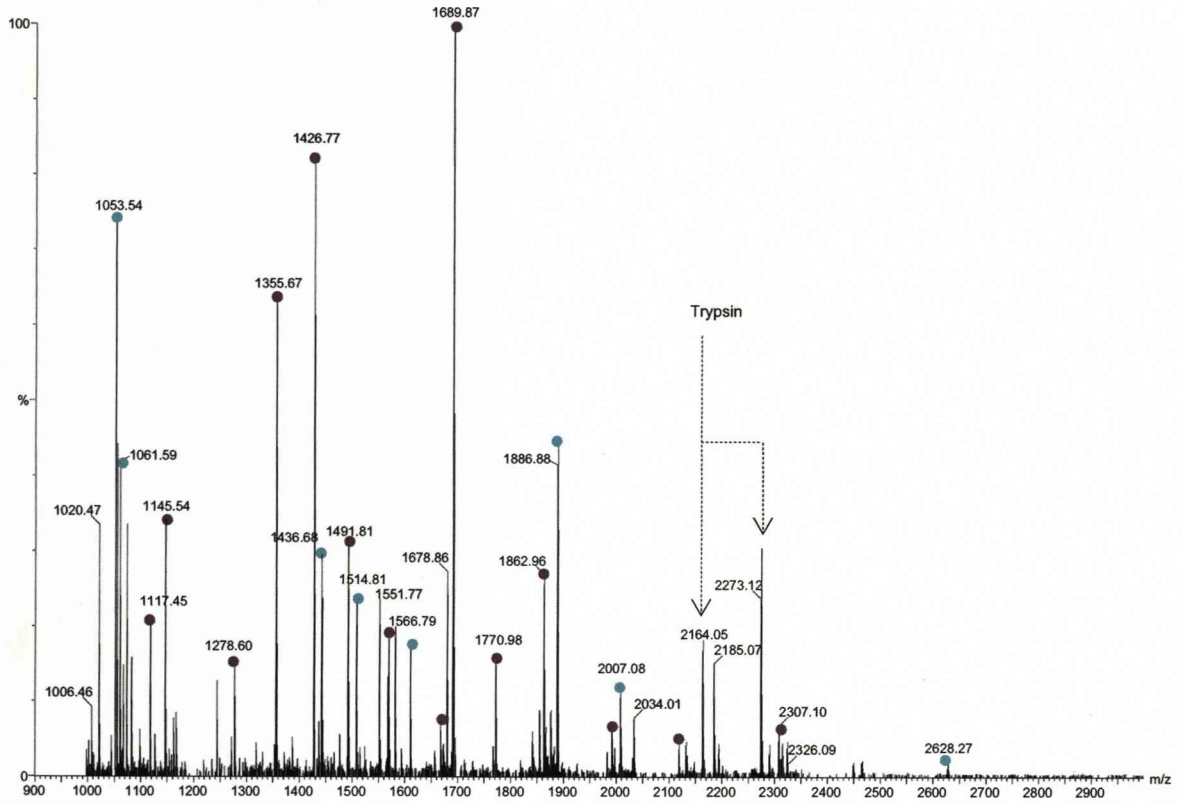
D. Identification of human plasma proteins by PMF

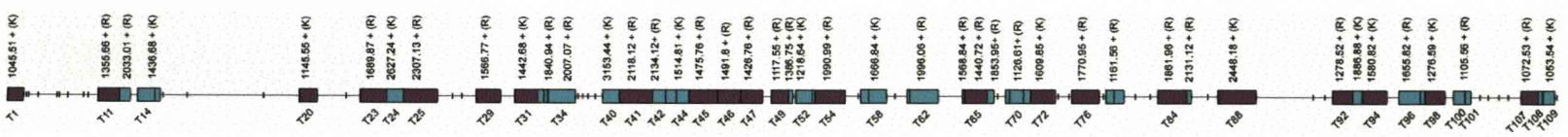
Supplementary Data A: Identification of mouse muscle proteins by PMF

- Glycogen phosphorylase
- Serum albumin
- Phosphoglucomutase-1
- Pyruvate kinase
- Beta enolase
- Creatine kinase
- Fructose-bisphosphate aldolase A
- Glyceraldehyde-3-phosphate dehydrogenase
- Phosphoglycerate mutase
- Triose phosphate isomerase
- Adenylate kinase

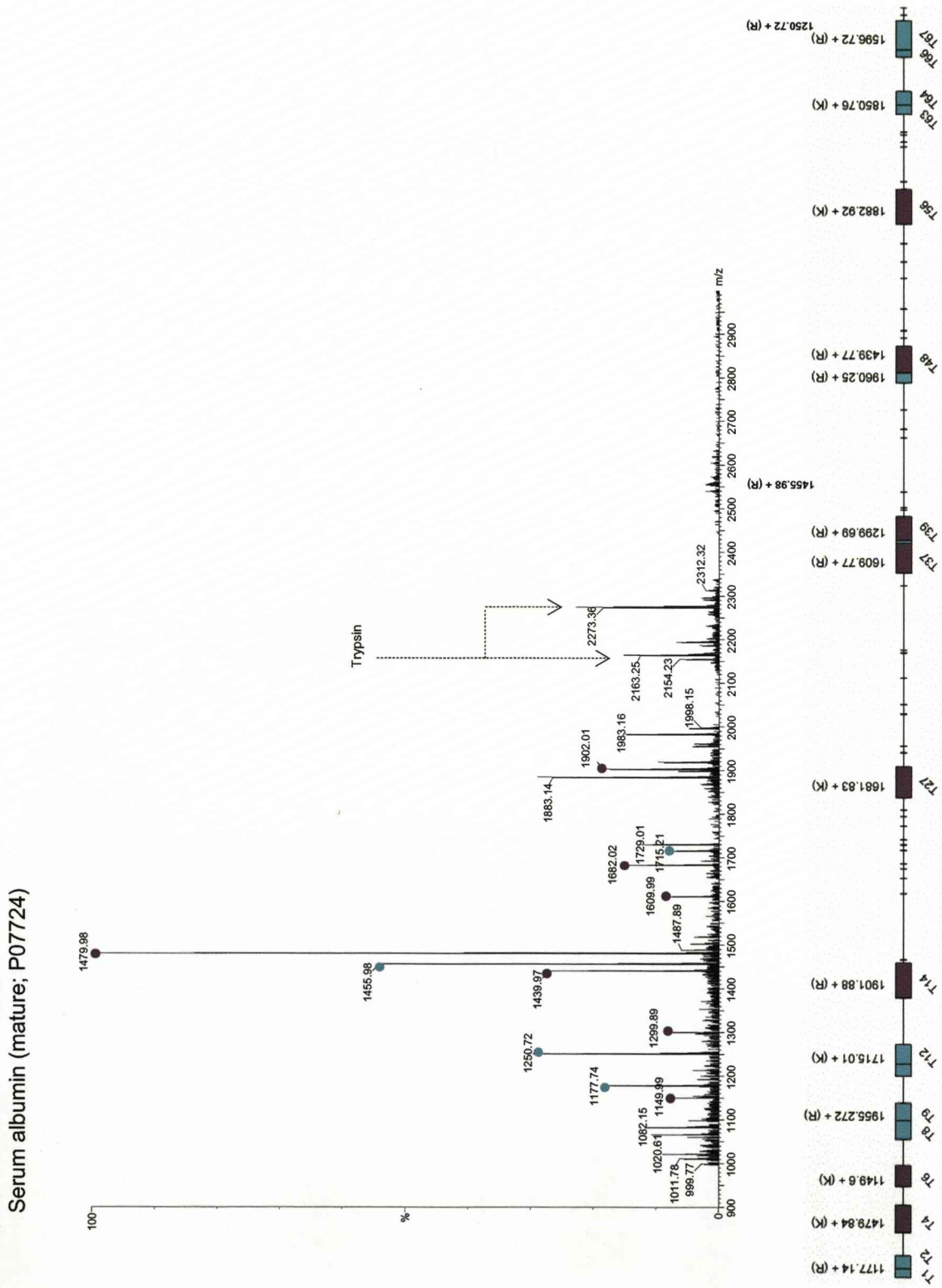
Proteins from mouse skeletal muscle soluble fraction (15µg) were separated by 1-D SDS-PAGE and visualised with Coomassie. Gel plugs were excised from the dominant bands and subjected to in-gel proteolysis with trypsin (1:50 enzyme to substrate ratio). Peptide mixtures 1µl were spotted onto a MALDI target and allowed to air dry with 1µl of matrix solution. Samples were analysed by MALDI-ToF MS using a laser energy of 30%. The resulting peptide masses were imported into the MASCOT search engine. The taxonomy was restricted to *Mus. musculus;* fixed modification: carbamidomethylation of cysteine; variable modification: oxidation of methionine; protease: trypsin; missed cleavages: 1; peptide tolerance: 150ppm. Matched peptides are represented on the peptide coverage maps. Limit tryptic peptides are represented in purple and missed cleavages in green.

Glycogen phosphorylase (Q9WUB3)

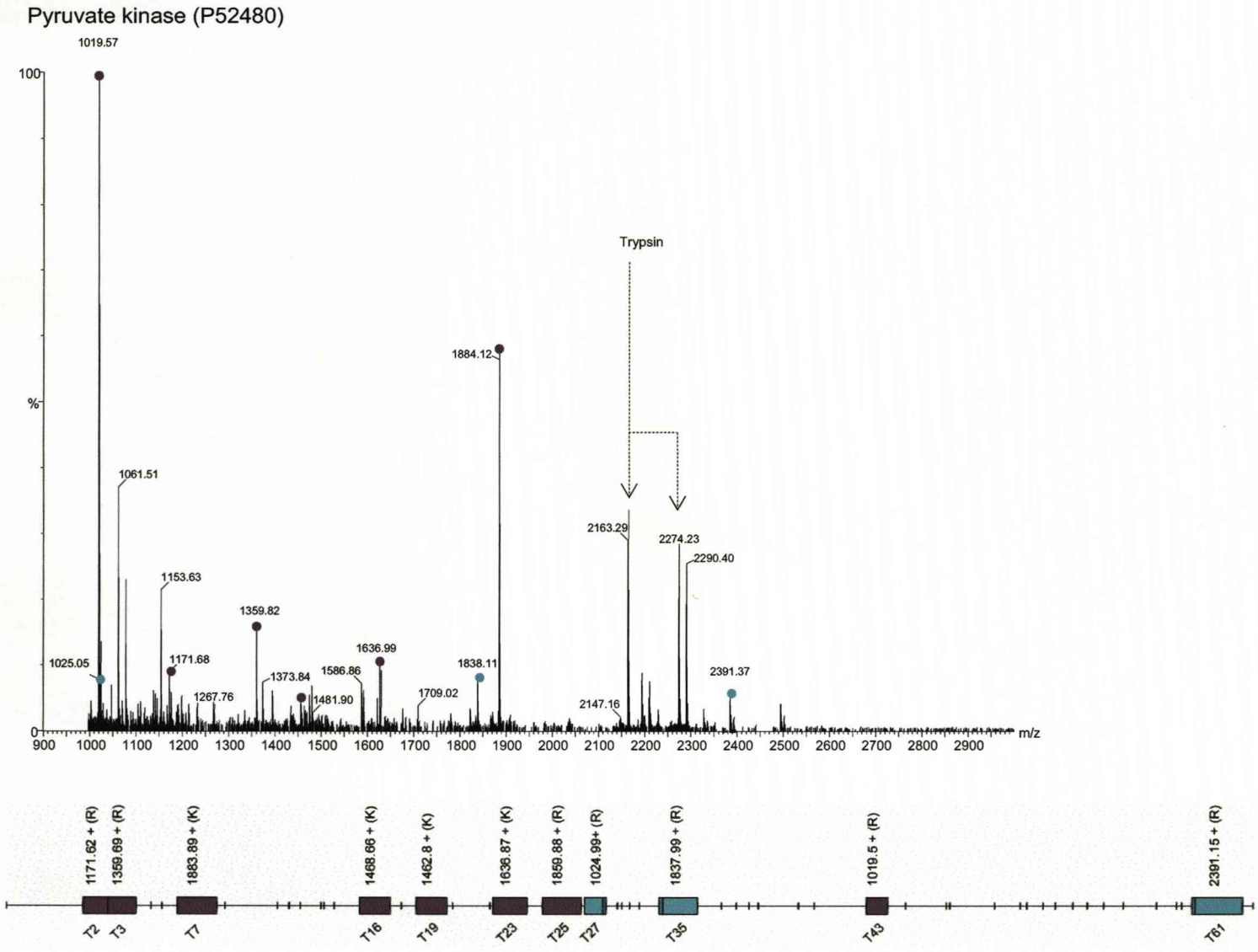




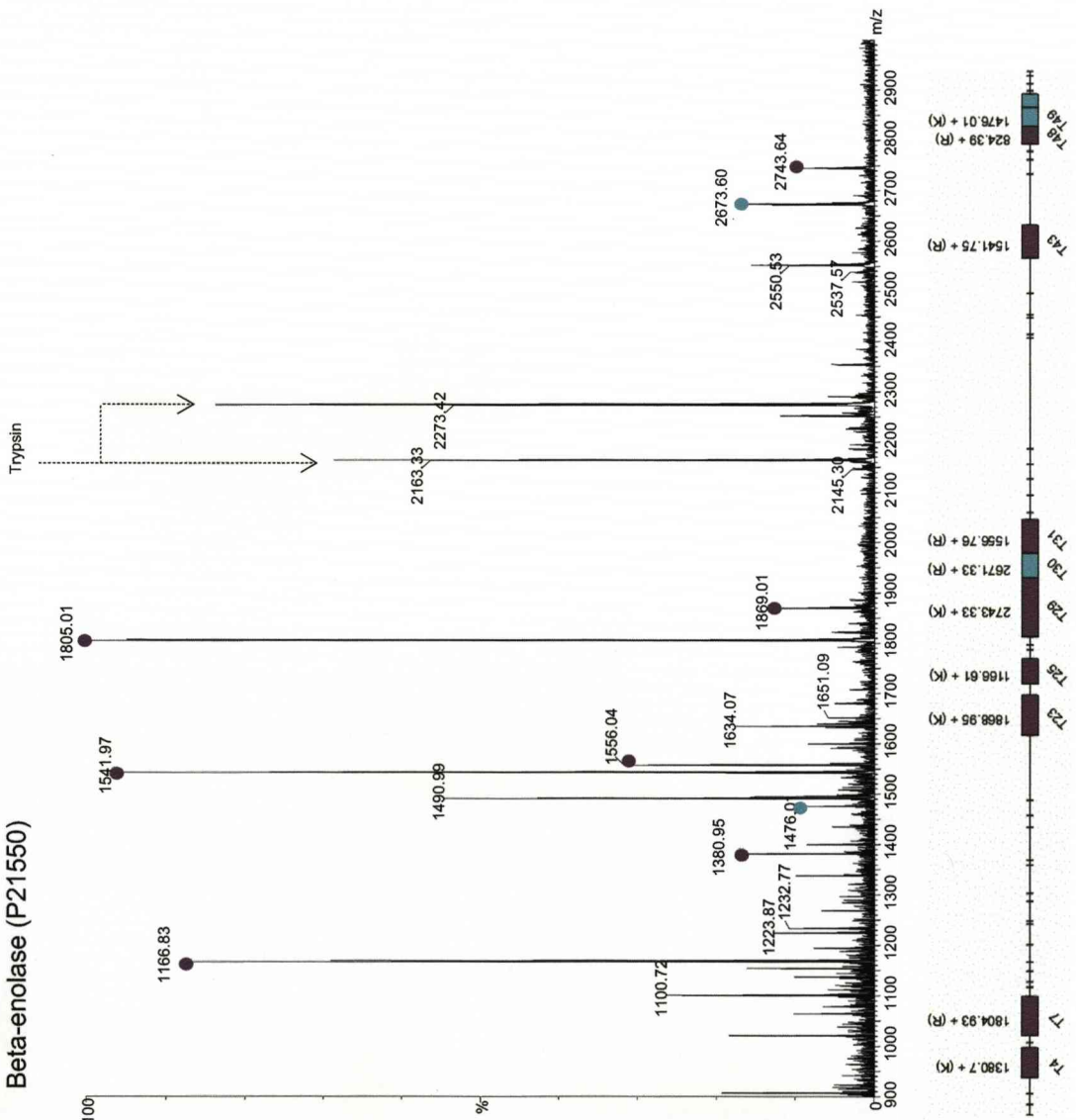
Supplementary data A: Identification of mouse muscle proteins by PMF


Serum albumin (mature; P07724)
Trypsin
m/z
%
1177.14 + (R)
1479.84 + (K)
1149.6 + (K)
1955.272 + (R)
1715.01 + (K)
1901.88 + (R)
1681.83 + (K)
1609.77 + (R)
1299.69 + (R)
1455.98 + (R)
1960.25 + (R)
1439.77 + (R)
1882.92 + (K)
1850.76 + (K)
1596.72 + (R)
1250.72 + (R)
T1 T2 T4 T6 T8 T9 T12 T14 T27 T37 T39 T48 T56 T63 T64 T66 T67

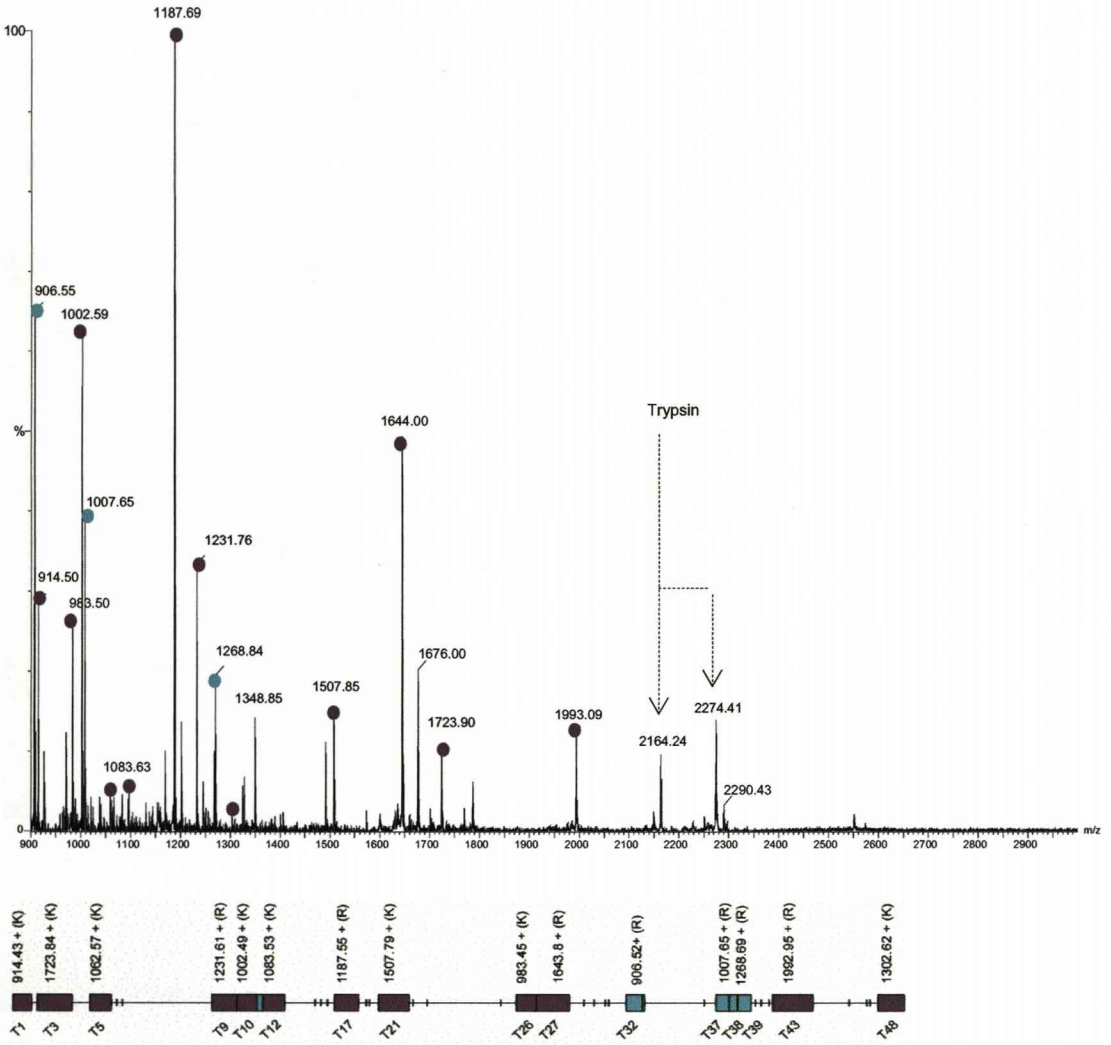
Phosphoglucomutase – 1 (Q9D0F9)
Trypsin
100
%
0
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
2400
2500
2600
2700
2800
2900
m/z
1081.65
1036.82
1020.58
1129.52
1091.63
1145.65
1153.69
1201.74
1278.82
1443.93
1385.90
1451.752
1484.02
1530.99
1619.06
1655.21
1754.19
1863.13
1879.15
1948.45
2011.47
2163.36
2273.53
2183.26
2206.15
2290.50
2326.33
2347.65
2515.75
2531.75
T3
1747.18+ (R)
T8
1201.6 + (R)
T9
1483.78 + (R)
T11
2011.12 + (R)
T14
2206.15 + (K)
T15
1630.79 + (K)
T17
2347.59+ (R)
1451.69 + (K)
T22
1081.55 + (K)
T23
1948.34 + (R)
T26
1655.14 + (K)
1385.71 + (K)
1401.91 + (K)
T29
1513.99+ (R)
2309.35+ (R)
T33
1619.06 + (R)
T35
1863.20+ (R)
T36
1278.66 + (K)
T37
1129.52 + (K)
1864.83 + (R)
2122.30+ (R)
T41
1626.07 + (K)
T45
2128.28 + (K)
T47
1075.53 + (R)
T49
1528.97 + (K)
T51
1743.75 + (K)
T56
1030.49 + (K)
1090.63+ (K)
1443.93 + (K)
1964.11 + (K)
T58



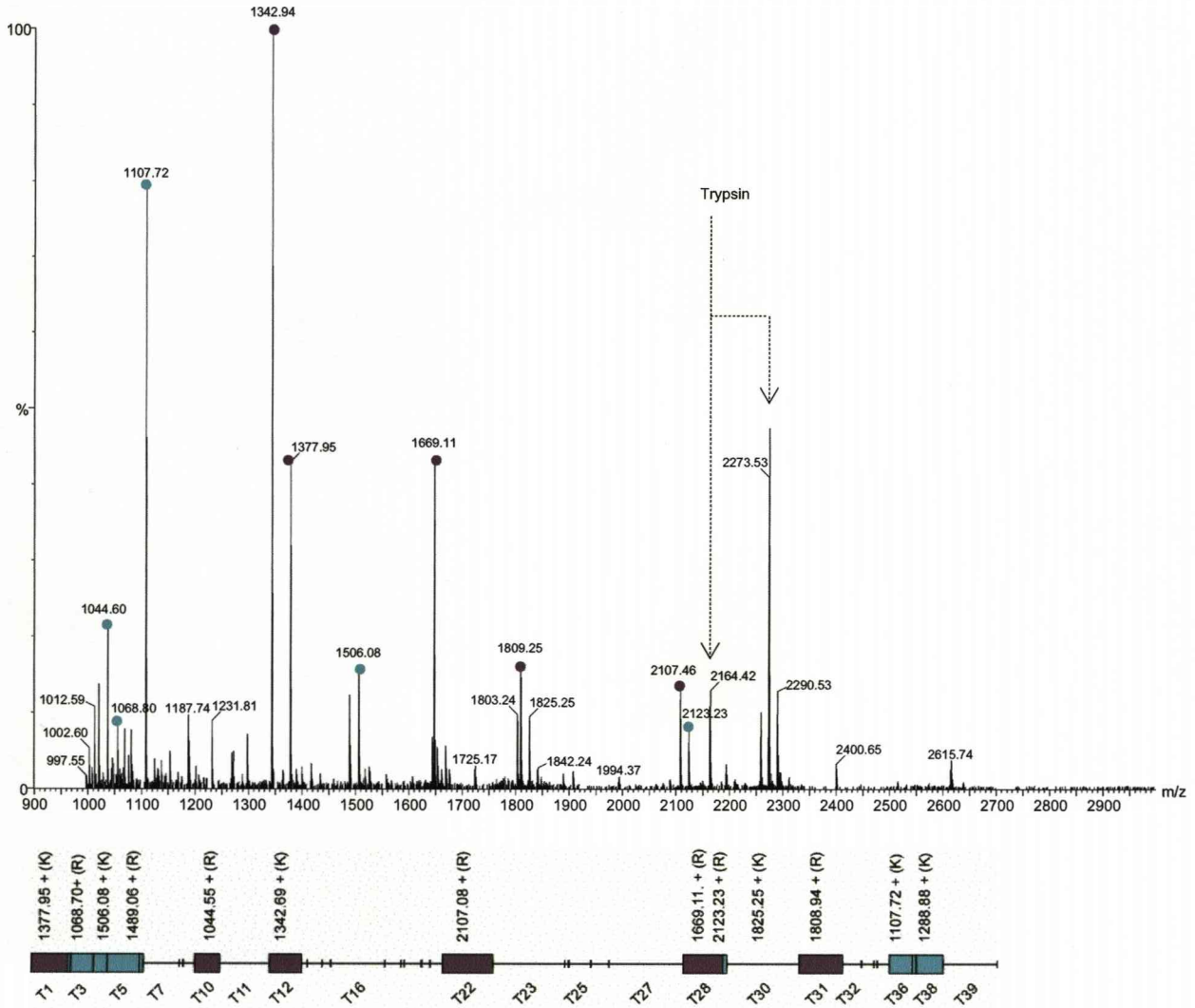
Supplementary data A: Identification of mouse muscle proteins by PMF


Beta-enolase (P21550)
Trypsin
1805.01
1541.97
1166.83
1490.99
1556.04
1634.07
1651.09
1869.01
2163.33
2273.42
2145.30
2550.53
2537.57
2673.60
2743.64
1100.72
1380.95
1476.01
1223.87
1232.77
m/z
%
100
0
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
2400
2500
2600
2700
2800
2900
T4
1380.7 + (K)
T7
1804.93 + (R)
T23
1868.95 + (K)
T25
1166.61 + (K)
T29
2743.33 + (K)
T30
2671.33 + (R)
T31
1556.76 + (R)
T43
1541.75 + (R)
T48/T49
824.39 + (R)
1476.01 + (K)

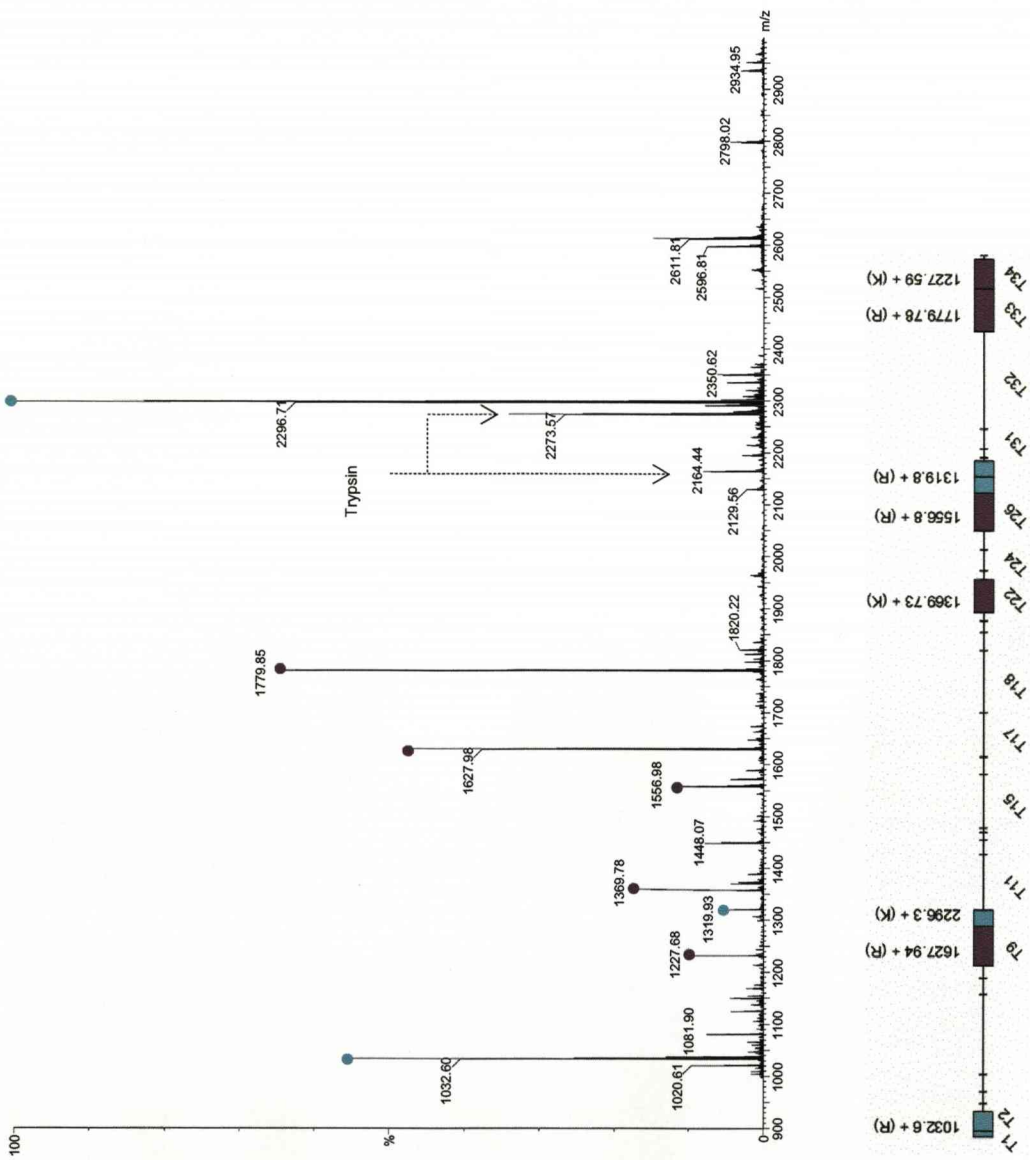
Creatine kinase (P07310)

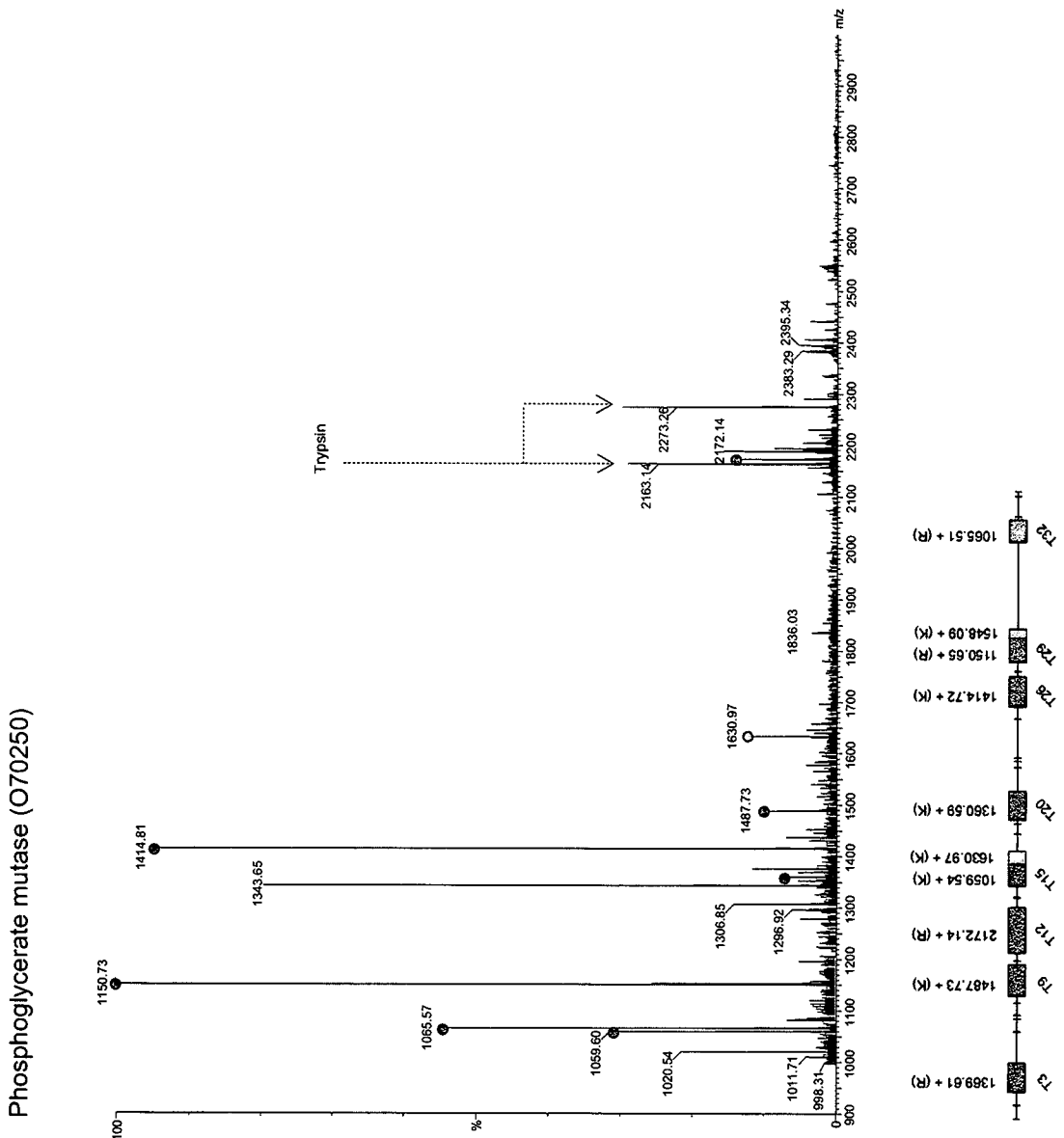


Fructose-bisphosphate aldolase A (P05064)

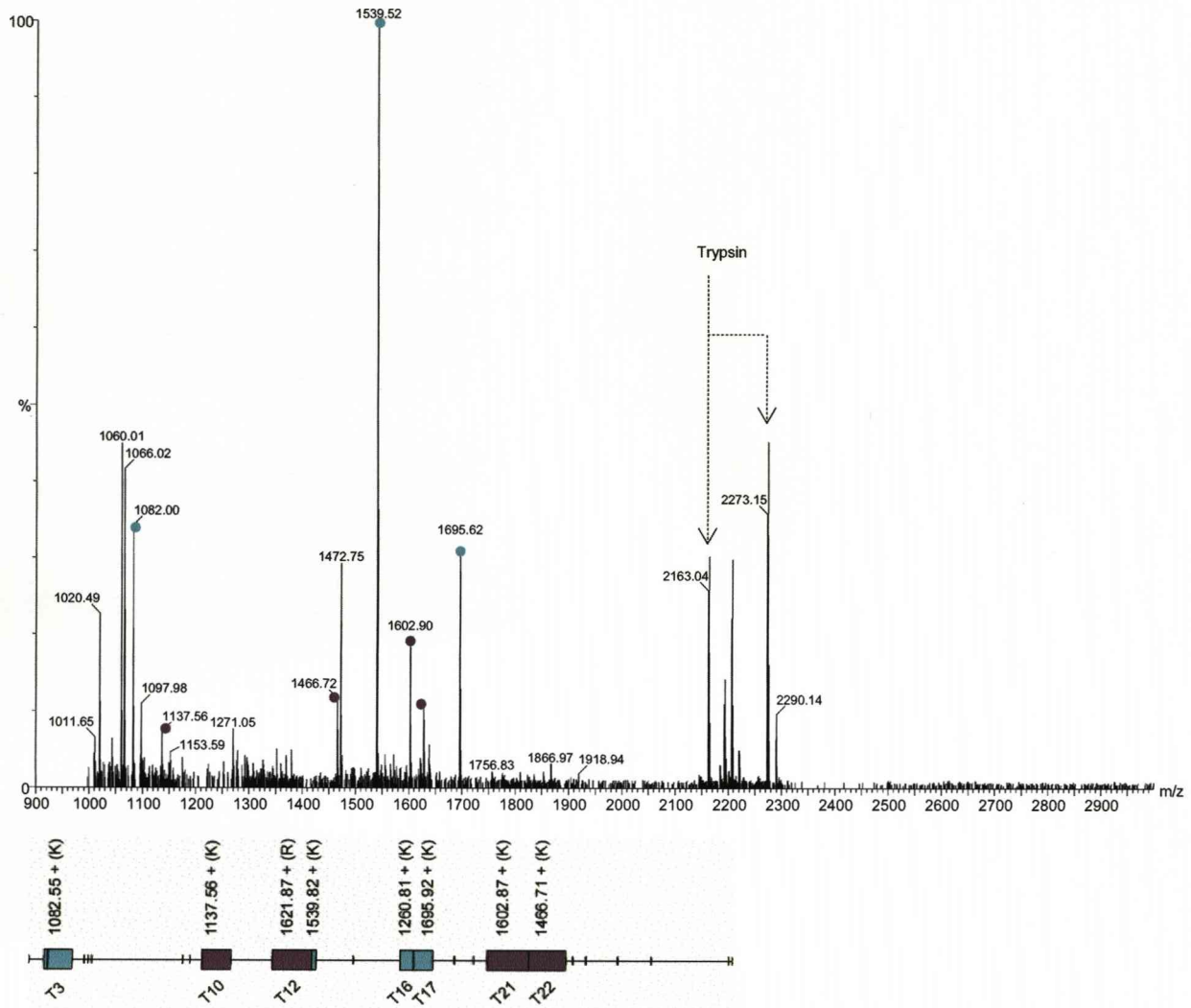


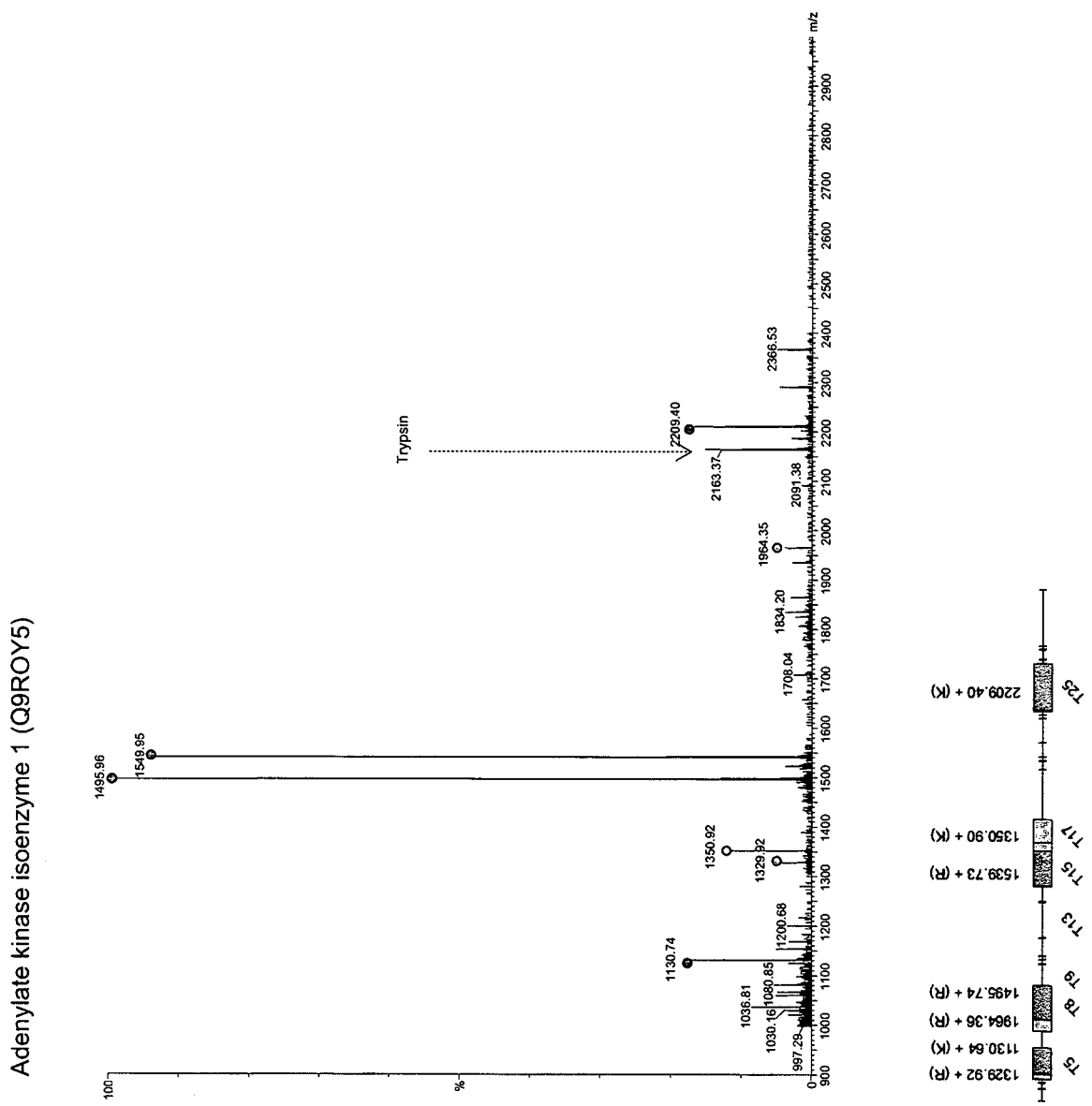
Glyceraldehyde-3-phosphate dehydrogenase (P16858)




Phosphoglycerate mutase (O70250)
Trypsin
m/z
%
1150.73
1414.81
1343.65
1065.57
1059.60
1020.54
1011.71
998.31
1487.73
1630.97
1836.03
1306.85
1296.92
2163.14
2273.26
2172.14
2383.29
2395.34
1369.61 + (R)
1487.73 + (K)
2172.14 + (R)
1059.54 + (K)
1630.97 + (K)
1360.58 + (K)
1414.72 + (K)
1150.65 + (R)
1548.09 + (K)
1065.51 + (R)

Triose phosphate isomerase (P17751)




Adenylate kinase isoenzyme 1 (Q9ROY5)
Trypsin
1495.96
1549.95
1130.74
1350.92
1329.92
1200.68
1708.04
1834.20
1964.35
2091.38
2209.40
2163.37
2366.53
997.29
1030.16
1036.81
1080.85
%
m/z
T25
2209.40 + (K)
T17
1350.90 + (K)
T15
1539.73 + (R)
T13
T9
1495.74 + (R)
T8
1964.36 + (R)
T3
1130.64 + (K)
1329.92 + (R)

Supplementary Data B: N-terminal identifications

- Mouse liver
- *S. cerevisiae*
- *E. coli*

The N-terminal peptide preparation of mouse skeletal muscle, *S. cerevisiae* and *E. coli* (soluble proteins) were analysed by LC-MS/MS using three hour RP gradients on the LTQ ion trap instrument. MS/MS data was used to search the mouse, *S. cerevisiae* and *E. coli* N-terminal databases respectively, using the MASCOT search engine. Fixed modifications: N-terminal acetylation and lysine acetylation; variable modification: oxidation of methionine; protease: Arg-C; missed cleavages: 1; peptide tolerance: 1.5Da, MS/MS tolerance: 0.6Da, instrument: ESI-TRAP, peptide charge: 1+, 2+ and 3+. The N-terminal processing events M (methionine excision) and SP (signal peptides removal) along with N^{α}-acetylation status are represented.

Supplementary data B: Mouse liver N-terminal identifications

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score	N^{α}-acetylation
1	Transcription factor Sp3 (O70494)	516.23	EGGGR	M	18	-
2	Golgi membrane protein 1 (Q91XA2)	614.31	GLGNGR	M	24	-
3	zinc finger domain protein 1B (Q9Z277)	667.4	APLLGR	M	12	N
4	Angiotensin-converting enzyme (P09470)	687.33	GAASGQR	M	14	-
5	Peroxisomal bifunctional enzyme (Q9DBM2)	692.35	AEYLR	M	20	-
6	Vitamin K epoxide reductase complex (Q6TEK5)	738.48	AAPVLLR	M	32	N
7	Alcohol dehydrogenase class 3 (P28474)	741.41	ANQVIR	M	32	N
8	BTB/POZ domain-containing protein (Q6WVG3)	747.4	ALADSAR	M	12	C
9	Adenylate kinase isoenzyme 4 (Q9WUR9)	770.47	ASKLLR	M	15	N
10	Glycine N-methyltransferase (Q9QXF8)	779.38	VDSVYR	M	89	N
11	Signal recognition particle 54 kDa protein (P14576)	784.44	VLADLGR	M	15	-
12	Translation initiation factor eIF-2B subunit delta (Q61749)	797.48	AAVAVAVR	M	14	N
13	T-complex protein 1 subunit delta (P80315)	813.4	PENVASR	M	52	N/C
14	Clathrin heavy chain (Q68FD5)	851.52	AQILPIR	M	27	N
15	Citrate synthase (Q9CZU6)	857.5	ALLTAATR	M	19	-
16	39S ribosomal protein L9 (Q99N94)	868.48	AASVAPGVR	M	147	N
17	Cysteine sulfinic acid decarboxylase (Q9DBE0)	872.48	ADSKPLR	M	36	C
18	Galactokinase (Q9R0N0)	894.48	AAWRPPR	M	28	N
19	Carbonic anhydrase 5B (Q9QZA0)	897.45	AVMNHLR	M	24	N
20	40S ribosomal protein S3a (P97351)	897.5	AVGKNKR	M	85	C
21	Quinone oxidoreductase (P47199)	897.5	ATGQKLMR	M	45	N
22	Ezrin (p81) (P26040)	906.53	PKPINVR	M	21	C
23	Ectodysplasin A receptor-associated adapter protein (Q8VHX2)	911.43	ASPDDPLR	M	101	-
24	Kynurenine--oxoglutarate transaminase 1 (Q8BTY1)	913.5	SKQLQAR	M	32	N
25	Guanine nucleotide-binding protein subunit beta (P68040)	919.44	TEQMTLR	M	53	-
26	Autocrine motility factor receptor (Q9R049)	928.54	PLLFLER	M	17	C
27	Bromodomain-containing protein 4 (Q9ESU6)	932.42	STESGPGTR	M	15	N
28	Glucose-6-phosphate 1-dehydrogenase (P97324)	932.47	AEQVTLSR	M	21	N
29	26S proteasome non-ATPase regulatory subunit 9 (Q9CR00)	937.43	SGEDVPHR	M	28	N
30	Inorganic pyrophosphatase (Q9D819)	939.39	SGFSSEER	M	67	N
31	60S ribosomal protein L39 (P62892)	945.47	SSHKTFR	M	58	N
32	60S ribosomal protein L18a (P62717)	946.49	MKASGTLR		107	-
33	40S ribosomal protein S30 (P62862)	950.53	KVHGSLAR	M	25	C
34	Thiosulfate sulfurtransferase (P52196)	955.52	VHQVLYR	M	74	C
35	Ankyrin repeat and SOCS box protein 1 (Q9WV74)	957.42	AEGGTGPDGR	M	21	N
36	TFIIH basal transcription factor complex p44 subunit (Q9JIB4)	962.37	MDEEPER		65	N
37	Glucosidase 2 subunit beta precursor (O08795)	966.56	VEVKRPR	SP	23	N
38	Eukaryotic initiation factor 4A-II (P10630)	967.4	SGGSADYNR	M	26	N
39	Protein C14orf4 homolog (Q8K3X4)	975.46	SAAQVSSSR	M	47	N
40	Ferritin heavy chain (P09528)	987.5	TTASPSQVR	M	36	-
41	40S ribosomal protein S24 (P62849)	990.48	MNDTVTIR		47	N
42	Fizzy-related protein homolog (Q9R1K5)	997.38	MDQDYER		23	N
43	Protein-glutamine gamma-glutamyltransferase 2 (P21981)	1013.54	AEELLLER	M	38	N
44	Fibroblast growth factor 14 (P70379)	1025.59	AAAIASGLIR	M	33	N
45	Mitochondrial 28S ribosomal protein S14 (Q9CR88)	1027.6	AASVLGSLLR	M	17	N
46	Nonspecific lipid-transfer protein (P32020)	1037.59	PSVALKSPR	M	45	C
47	Embryonal Fyn-associated substrate (Q64355)	1042.58	AIATSAQLAR	M	46	-
48	40S ribosomal protein S3 (P62908)	1054.65	AVQISKKR	M	103	N
49	Thiamine-triphosphatase (Q8JZL3)	1055.56	AQGLIEVER	M	20	-
50	Protein patched homolog 1 (Q61115)	1056.53	ASAGNAAGALGR	M	26	N
51	Calcium-binding protein 1 (CaBP1) (Q9JLK7)	1056.54	GNCVKSPLR	M	10	N
52	Nephrin (Q9QZS7)	1080.56	QSPVPTSAPR	SP	19	-
53	3'(2'),5'-bisphosphate nucleotidase 1	1084.53	ASSHTVLMR	M	45	N
54	Succinate dehydrogenase (Q9CQA3)	1084.62	AATVGVSLKR	M	12	N
55	Succinyl-CoA ligase [GDP-forming] subunit alpha (Q9WUM5)	1088.55	VSSSSGLAAAR	M	33	-
56	SON protein (Q9QX47)	1089.55	AADIEQVFR	M	177	N
57	ADAMTS-15 (P59384)	1098.9	AIPAGASSIDR	M	15	N
58	Esterase D (Q9R0P3)	1099.6	ALKQISSNR	M	81	N
59	5-formyltetrahydrofolate cyclo-ligase (Q9D110)	1099.6	AAVTVNSAKR	M	39	-
60	Long-chain-fatty-acid--CoA ligase 1	1101.53	MEVHELFR		12	C
61	Gastrin/cholecystokinin type B receptor (P56481)	1101.59	MDLLKLNR		47	N
62	26S proteasome regulatory subunit S9 (Q8BG32)	1102.58	AAAAVVEFQR	M	67	N
63	NG,NG-dimethylarginine dimethylaminohydrolase 1 (Q9CWS0)	1110.56	AGLGHPSAFGR	M	42	N
64	Glyceraldehyde-3-phosphate dehydrogenase (P16858)	1115.61	VKVGVNGFGR	M	118	C
65	Adenylyl cyclase-associated protein 1 (P40124)	1116.52	ADMQNLVER	M	31	-
66	Long-chain-fatty-acid--CoA ligase 1 (P41216)	1117.52	MEVHELFR		29	N
67	M-phase inducer phosphatase 1 (P48964)	1121.55	MELGPSPPPR		18	N
68	Elongation factor 2 (P58252)	1132.59	VNFTVDQIR	M	58	C
69	Acyl-coenzyme A thioesterase 4 (Q8BWN8)	1142.59	AATLSVEPTGR	M	17	N
70	40S ribosomal protein S10 (P63325)	1142.59	MLMPKKNR		65	N

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score	N^{α}-acetylation
71	Polymerase delta-interacting protein 3 (Q8BG81)	1143.61	ADLSLDELIR	M	45	C
72	High mobility group protein 1 (P63158)	1149.61	GKGDPKKPR	M	38	N
73	14-3-3 protein eta (P68510)	1155.6	GDREQLLQR	M	39	N
74	Estradiol 17 beta-dehydrogenase 5 (P70694)	1175.56	MDSKQQTVR		125	N
75	Importin alpha-3 subunit (O35344)	1177.55	AENPGLENHR	M	12	N
76	Hippocampus abundant transcript 1 protein (P70187)	1182.67	TQGKKKKR	M	15	-
77	Annexin A6	1191.61	AKIAQGAMYR	M	152	N
78	Ezrin-radixin-moesin binding phosphoprotein 50 (P70441)	1195.58	SADAAAGEPLPR	M	33	N
79	Oligodendrocyte transcription factor 3 (Q6PFG8)	1197.49	MNSDSSSVSSR		15	N
80	60S ribosomal protein L13a (P19253)	1197.64	AEGQVLVLDGR	M	70	N
81	5'-AMP-activated protein kinase catalytic subunit alpha-1 (Q5EG47)	1199.62	AEKQKHDGR	M	14	N
82	Ras association domain-containing protein 1 (Q99MK9)	1200.64	SAEPELIELR	M	11	C
83	2 days neonate thymus thymic cells cDNA (Q8C7C3)	1229.66	AGTTTIEAVKR	M	15	N
84	Argininosuccinate lyase (Q91YI0)	1230.6	ASESGKLWGGR	M	158	N
85	Tropomyosin alpha-4 chain (Q6IRU2)	1240.68	AGLNSLEAVKR	M	27	N
86	Selenoprotein T precursor (P62342)	1252.64	SANLGGVPSK*R	SP	48	N
87	Aldehyde dehydrogenase (P47739)	1257.68	SNISSIVNRAR	M	65	N
88	Serum albumin precursor (P07724)	1260.62	EAHKSEIAHR	PP	129	C
89	Phenylalanine-4-hydroxylase (P16331)	1268.71	AAVVLENGVLSR	M	126	N
90	Laminin subunit alpha-3 mature (Q61789)	1269.58	VVGQDHPMSSR	M	19	-
91	Fatty acid-binding protein (P55050)	1277.6	AFDGTWKVDR	M	92	N
92	40S ribosomal protein S15 (P62843)	1282.69	AEVEQKKKR	M	79	N
93	Heterogeneous nuclear ribonucleoproteins C1/C2 (Q9Z204)	1285.63	ASNVTNKTDPR	M	49	N
94	Glutathione S-transferase Mu 1 (P10649)	1289.66	PMILGYWNVR	M	247	C
95	Glutathione S-transferase Mu 2 (P15626)	1292.62	PMTLGYWDIR	M	102	C
96	JmjC domain-containing histone demethylation protein 3A (Q8BW72)	1302.61	ASESETLNPSAR	M	25	N
97	Glycogen phosphorylase, liver form (Q9ET01)	1310.68	AKPLTDQEKR	M	105	N
98	Dihydropyrimidinase-related protein 3 (Q62188)	1315.69	SYQGKKNIPR	M	17	N
99	Paired box protein Pax-5 (Q02650)	1318.66	DLEKNYPTPR	M	158	N
100	Homeobox protein orthopedia (O09113)	1324.64	MLSHADLLDAR		123	-
101	Macrophage migration inhibitory factor (P34884)	1331.71	PMFIVNTNVPR	M	208	C
102	Peptidyl-prolyl cis-trans isomerase (P26883)	1355.67	GVQVETISPGDGR	M	45	N
103	Tyrosyl-tRNA synthetase (Q91WQ3)	1355.67	GDAPSPEEKLHLITR	M	36	-
104	Acyl-coenzyme A thioesterase 1 (O55137)	1358.65	MEATLNLEPSGR		28	N
105	Exosome complex exonuclease RRP41 (Q921I9)	1362.68	AGLELLSDQGYR	M	22	N
106	Sterol-4-alpha-carboxylate 3-dehydrogenase (Q9R1J0)	1370.64	MEQAVHGESKR		20	N
107	Glutathione S-transferase P 1 (P19157)	1392.74	PPYTIVYFPVR	M	356	C
108	Eukaryotic translation initiation factor 1A (Q8BMJ3)	1392.75	PKNKGKGGKNR	M	117	C
109	Eukaryotic translation initiation factor 3 subunit I (Q9QZD9)	1404.75	MKPILLQGHER		98	-
110	Mitochondrial import receptor subunit TOM34 (Q9CYG7)	1411.73	APKVSDSVEQLR	M	27	C
111	14-3-3 protein gamma (P61982)	1424.77	VDREQLVQKAR	M	31	N
112	Fatty acid-binding protein, epidermal (E-FABP) (Q05816)	1427.33	ASLKDLEGKWR	M	29	N
113	Fatty acid-binding protein (Q05816)	1427.74	ASLKDLEGKWR	M	58	N
114	Male-enhanced antigen 1 (Q64327)	1429.69	AAVVLGGDTMGPER	M	26	N
115	Histone deacetylase 5 (Q9Z2V6)	1438.55	MNSPNESDGMSGR		16	-
116	Septin-11 (Q8C1B7)	1438.75	AVAVGRPSNEELR	M	30	N
117	Glutathione S-transferase Yc (P30115)	1442.73	AGKPVLHYFDGR	M	161	N
118	Mitochondrial import inner membrane translocase (P62077)	1471.68	AELGEADEAELQR	M	75	N
119	60S ribosomal protein L11 (Q9CXW4)	1485.65	AQDQGEKENPMR	M	50	N
120	Sal-like protein 3 (Q62255)	1493.59	PGDGAEDADSGSESR	M	45	C
121	Signal recognition particle 9 kDa protein (P49962)	1495.67	PQFQTWEEFSR	M	15	C
122	D-dopachrome tautomerase (O35215)	1513.78	PFVELETNLPASR	M	349	C
123	Cytoplasmic dynein 1 intermediate chain 1(O88485)	1515.74	SDKSDLKAELER	M	10	N
124	4-hydroxyphenylpyruvate dioxygenase	1529.75	TTYNNKGPKPER	M	75	-
125	L-serine dehydratase (Q8VBT2)	1532.83	AAQESLHVKTPLR	M	128	N
126	Glutathione S-transferase P 2 (P46425)	1534.78	PPYTIVYFPSPGR	M	50	N
127	Peroxisomal carnitine O-octanoyltransferase (Q9DC50)	1546.73	MENQLTKSVEER		20	N
128	Polypyrimidine tract-binding protein 1 (P17225)	1554.81	MDGIVPDIAVGTKR		45	N
129	Ras suppressor protein 1 (Q01730)	1570.86	SKSLKKLVEESR	M	58	N
130	Eukaryotic translation initiation factor 2 subunit 3 (Q9Z0N1)	1576.8	AGGEGGVTLGQPHLSR	M	40	N
131	Serine/threonine-protein phosphatase 6 (Q9CQR6)	1586.84	APLDLDKYVEIAR	M	22	-
132	Kynureninase (Q9CXF0)	1593.81	MEPSPLELPVDAVR		83	N
133	T-complex protein 1 subunit zeta (P80317)	1598.93	AAVKTLNPKAEVAR	M	59	N
134	Microsomal triglyceride transfer protein large subunit (O08601)	1608.82	VKGHTTGLSLNNER	M	39	-
135	Liver carboxylesterase 31 precursor (Q63880)	1619.85	PKVTQPEVDTPLGR	SP	100	C
136	Serine/threonine-protein phosphatase PP1 (P62137)	1629.82	SDSEKLNLDSIIGR	M	35	-
137	60S ribosomal protein L29 (P47915)	1633.79	AKSKNHTTHNQSR	M	25	-
138	Heterogeneous nuclear ribonucleoprotein A1 (P49312)	1639.81	SKSESPKEPEQLR	M	41	N
139	Hnrpa3 protein	1648.75	MEGHDPKEPEQLR		25	N
140	Vitamin K-dependent protein Z precursor (Q9CQW3)	1651.87	SVFLPAPKANNVLR	SP	48	N

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score	N^{α}-acetylation
141	Sorting nexin-5 (Q9D8U8)	1667.8	AAVPELLEQQEEDR	M	73	N
142	C-1-tetrahydrofolate synthase, cytoplasmic (Q922D8)	1690.97	APAGILNGKLVSAQIR	M	38	C
143	Protein disulfide-isomerase A3 precursor (P27773)	1694.76	DVLELTDENFESR	SP	32	-
144	Protein disulfide-isomerase A3 precursor (P27773)	1694.76	SDVLELTDENFESR	SP	183	C
145	Protein transport protein Sec23A (Q01405)	1711.81	TTYLEFIQQNEER	M	120	-
146	Fructose-1,6-bisphosphatase 1 (Q9QXD6)	1713.83	ANHAPFETDISTLTR	M	304	N
147	Signal recognition particle 14 kDa protein (P16254)	1718.91	VLLESEQFLTELTR	M	59	-
148	Acyl-CoA-binding protein (P31786)	1732.83	SQAEFDKAAEEKVR	M	391	N
149	F-actin-capping protein subunit alpha-2 (P47754)	1743.82	ADLEEQLSDEEKVR	M	110	-
150	Serine/threonine-protein kinase PRP4 homolog (Q61136)	1754.9	AAGIGKDFKENPNLR	M	98	-
151	Liver fructose 1, 6 Bisphoshate (O-accetylated) (Q9QXD6)	1755.83	ANHAPFETDIS*TLTR	M	45	-
152	BTG3 protein (Tob5 protein) (P50615)	1755.91	MKNEIAAVVFFFTR		23	C
153	40 kDa peptidyl-prolyl cis-trans isomerase (Q9CR16)	1773.88	SHASPAAKPSNSKNPR	M	95	-
154	Adenylate kinase isoenzyme 2 (Q9WTP6)	1773.96	APNVLASEPEIPKGIR	M	22	C
155	Betaine--homocysteine S-methyltransferase (O35490)	1775.03	APVAGKKAKKGILER	M	678	C
156	40S ribosomal protein S19 (Q9CZX8)	1798.92	PGVTVKDVNQQEFVR	M	189	C
157	Docking protein 1 (P97465)	1819.86	MDGAVMEGPLFLQSQR		16	-
158	Glutamate–cysteine ligase catalytic subunit (P97494)	1828.91	GLLSQGSPLSWEETQR	M	77	C
159	Splicing factor 3B subunit 1 (Q99NB9)	1847.97	AKIAKTHEDIEAQIR	M	105	N
160	Developmentally-regulated GTP-binding protein 2 (Q9QXB9)	1853.01	GILEKISEIEKEIAR	M	38	-
161	Guanine nucleotide-binding protein (Q61017)	1895.04	MEVEQLKKEVKNPR		28	N
162	60S ribosomal protein L32 (P17932)	1926.21	AALRPLVKPKIVKKR	M	70	-
163	60S ribosomal protein L30 (P62889)	1969.08	VAAKKTKKSLESINSR	M	89	-
164	Cytosolic nonspecific dipeptidase (Q9D1A2)	1979.96	SALKAVFQYIDENQDR	M	67	N
165	40S ribosomal protein S27	2026.04	PLAKDLLHPSPEEEKR	M	103	C
166	26S proteasome non-ATPase regulatory subunit 6 (Q99JI4)	2046.03	PLENLEEEGLPKNPDLR	M	29	C
167	Peptidyl-prolyl cis-trans isomerase A (P17742)	2046.99	VNPTVFFDITADDEPLGR	M	243	N/C
168	Guanidinoacetate methyltransferase (O35969)	2046.99	SSSAASPLFAPGEDCGPAWR	M	44	N
169	Endoplasmin mature (P08113)	2060.9	DDEVDVDGTVEEDLGKSR	SP	81	-
170	Histone H2A.Z (H2A/z)(P0C0S6)	2068.09	AGGKAGKDSGKAKTKAVSR	M	18	N
171	T-complex protein 1 subunit beta (P80314)	2071.06	ASLSLAPVNIFKAGADEER	M	51	N
172	Kruppel-like factor 2 (Q60843)	2077.02	ALSEPILPSFATFASPCER	M	46	N
173	Thyrotroph embryonic factor (Q9JLC6)	2081.99	SDAGGGKKPPVEPQAGPGPGR	M	58	N
174	PCTP-like protein (PCTP-L) (Q9JMD3)	2082.02	MEKPAASTEPQGSRPALGR		35	N
175	Sorbitol dehydrogenase	2084.1	AAPAKGENLSLVVHGPGDIR	M	85	N
176	Adenosylhomocysteinase (P50247)	2085.09	SDKLPYKVADIGLAAWGR	M	92	N
177	40S ribosomal protein S20 (P60867)	2120.09	AFKDTGKTPVEPEVAIHR	M	119	N
178	Phosphoglycerate kinase 1 (P09411)	2124.18	SLSNKLTLDKLDVKGKR	M	98	N
179	Serine/threonine-protein phosphatase 2A (Q76MZ3)	2158.08	AAADGDDSLYPIAVLIDELR	M	32	N
180	myosin regulatory light chain	2166.21	SSKKAKTKTTKKRPQR	M	47	N
181	26S protease regulatory subunit 8 (P62196)	2170.02	ALDGPEQMELEEGKAGSGLR	M	43	N
182	Peroxiredoxin 6 (O08709)	2183.06	PGGLLLGDEAPNFEANTTIGR	M	247	C
183	Homogentisate 1,2-dioxygenase	2210.98	AELKYISGFGNECASEDP	M	58	-
184	Cofilin-1 (P18760)	2233.14	ASGVAVSDGVIKVFNDMKVR	M	128	N
185	Abhydrolase domain-containing protein 14B (Q8VCR7)	2285.12	AGVDQHEGTIQVQGQNLFFR	M	64	N
186	14-3-3 protein theta (P68254)	2292.23	MEKTELIQKAKLAEQAER		321	N
187	14-3-3 protein zeta/delta (P63101)	2293.19	MDKNELVQKAKLAEQAER		383	N
188	Arginase-1 (Q61176)	2294.23	SSKPKSLEIIGAPFSKGQPR	M	24	N
189	Small nuclear ribonucleoprotein Sm D2 (P62317)	2295.18	SLLNKPKSEMTPEELQKR	M	21	-
190	Catalase (P24270)	2317.04	SDSRDPASDQMKQWKEQR	M	78	N
191	Cytoplasmic aconitate hydratase (P28271)	2332.16	MKNPFAHLAEPLDAAQPGKR		29	-
192	14-3-3 protein beta/alpha (Q9CQV8)	2341.25	TMDKSELVQKAKLAEQAER	M	125	-
193	Cofilin-2 (P45591)	2346.19	ASGVTVNDEVIKVFNDMKVR	M	38	N
194	14-3-3 protein epsilon (P62259)	2363.11	MDDREDLVYQAKLAEQAER		66	N
195	Calreticulin (P14211)	2369.1	DPAIYFKEQFLDGDAWTNR	SP	109	-
196	Retinitis pigmentosa 1-like 1 protein (Q8CGM2)	2424.09	MNSTPGDTRDAPAPSHPAPSHR		26	-
197	Selenium-binding protein 2 (Q63836)	2473.18	ATKCTKCGPGYPTPLEAMKGPR	M	26	N
198	Histidine triad nucleotide-binding protein 1 (P70349)	2523.33	ADEIAKAQVAQPGGDTIFGKIIR	M	58	N
199	Phosphoglucomutase-1 (Q9D0F9)	2553.38	VKIVTVKTQAYPDQKPGTSGLR	M	28	-
200	Fructose-bisphosphate aldolase B (Q91Y97)	2580.39	AHRFPALTPEQKKELSEIAQR	M	100	N
201	Transitional endoplasmic reticulum ATPase(Q01853)	2709.37	ASGADSKGDDLSTAILKQKNRPNR	M	29	N
202	Thioredoxin-like protein 2 (Q9CQM9)	2784.39	AAGAAEAGEAAVAVVEVGSAQQFE	M	39	N
203	Lysyl-tRNA synthetase (Q99MN1)	2838.46	ATLQESEVKVDGEQKLSKNELKR	M	41	N
204	Glutathione synthetase (P51855)	2854.43	ATSWGSILQDEKQLEELAKQAIDR	M	32	N
205	U6 snRNA-associated Sm-like protein LSm7 (Q9CQQ8)	3084.66	ADKEKKKKESILDLSKYIDKTIR	M	37	-
206	Hemoglobin beta-1 subunit (P02088)	3278.59	VHLTDAEKAAVSCLWGKVNSDEVG	M	15	-
207	40S ribosomal protein S12 (P63323)	3361.72	AEEGIAAGGVMDVNTALQEVLKTALI	M	71	N
208	Hemoglobin alpha subunit (P01942)	3380.66	VLSGEDKSNIKAAWGKIGGHGAEYG	M	337	C
209	Alpha-1-antitrypsin 1-4 (Q00897)	3385.55	EDVQETDTSQKDQSPASHEIATNLG	SP	40	-
210	Aminopeptidase B (Q8VCT3)	3400.59	MESGGPGNYSGAARRPLHSAQAVD		25	

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score	N^{α}-acetylation
1	40S ribosomal protein S19-B (P07281)	629.35	AGVSVR	M	45	C
2	40S ribosomal protein S19-A (P07280)	655.37	PGVSVR	M	26	C
3	40S ribosomal protein S23 (P32827)	767.43	GKGKPR	M	28	C
4	40S ribosomal protein S24 (P26782)	802.42	SDAVTIR	M	48	N
5	40S ribosomal protein S14-A (P06367)	814.43	SNVVQAR	M	36	N
6	APC/C-CDH1 modulator 1 (Q08981)	827.45	SPSKKR	M	14	N
7	Vacuolar ATP synthase catalytic subunit A (P17255)	842.42	AGALENAR	M	19	N
8	60S ribosomal protein L39 (P04650)	848.45	AAQKSFR	M	65	C
9	60S ribosomal protein L42 (P02405)	896.51	VNVPKTR	M	78	C
10	40S ribosomal protein S1-A (P23248)	897.50	AVGKNKR	M	47	C
11	Phosphoglycerate mutase 1 (P00950)	907.59	PKLVLVR	M	147	C
12	Farnesyl pyrophosphate synthetase (P08524)	915.47	ASEKEIR	M	101	-
13	40S ribosomal protein S14-B (P39516)	927.48	ANDLVQAR	M	169	C
14	Flavohemoprotein (P39676)	931.48	MLAEKTR		29	C
15	Adenylate kinase cytosolic (P07170)	932.41	SSSESIR	M	57	N
16	TPR repeat-containing protein YDR161W (Q03771)	959.49	SELEATIR	M	14	N
17	Threonine synthase (P16120)	975.48	PNASQVYR	M	65	C
18	Enolase 1 (P00925)	976.53	AVSKVYAR	M	98	C
19	UPF0202 protein YNL132W (P53914)	1013.55	AKKAIDSR	M	21	C
20	40S ribosomal protein S30 (Q12087)	1021.57	AKVHGSLAR	M	59	N
21	Fructose-bisphosphate aldolase (P14540)	1025.59	GVEQILKR	M	88	C
22	40S ribosomal protein S3 (P05750)	1039.34	VALISKKR	M	36	C
23	60S ribosomal protein L31 (P0C2H9)	1041.27	AGLKDVVTR	M	67	C
24	Guanine nucleotide-binding protein (P38011)	1041.58	ASNEVLVLR	M	54	N
25	60S ribosomal protein L11-A (P0C0W9)	1043.52	SAKAQNPMR	M	47	N
26	PAB-dependent poly(A)-specific ribonuclease subunit PAN3 (P36102)	1048.66	KLLKYPR	M	12	N
27	Putative uncharacterized protein YIL058W (P40521)	1053.50	MDGHVQGLR		19	N
28	Translation machinery-associated protein 20 (P89886)	1082.56	MFKKFTR		8	-
29	Fimbrin (P32599)	1084.61	MNIVKLQR		47	N
30	3-isopropylmalate dehydratase (P07264)	1087.57	VYTPSKGPR	M	24	-
31	60S ribosomal protein L21-A (Q02753)	1088.55	GKSHGYRSR	M	69	C
32	60S ribosomal protein L1 (P53030)	1088.58	SKITSSQVR	M	84	N
33	60S ribosomal protein L35 (P39741)	1089.58	AGVKAYELR	M	87	N/C
34	Methylthioribose-1-phosphate isomerase (Q06489)	1090.41	SLEAIVFDR	M	12	N
35	Elongation factor 1-gamma 2 (P36008)	1092.56	SQGTLYINR	M	65	N
36	40S ribosomal protein S12 (P48589)	1093.59	TALVHDGLAR	M	47	-
37	40S ribosomal protein S10-A (Q08745)	1102.52	MLMPKEDR	M	69	N
38	Elongation factor 2 (EF-2) (P32324)	1107.54	VAFTVDQMR	M	54	C
39	Orotidine 5'-phosphate decarboxylase (P03962)	1107.56	SKATYKER	M	12	N
40	Rho GDP-dissociation inhibitor (Q12434)	1113.19	AEESTDFSQFEEER	M	32	N
41	Uncharacterized protein YMR074C (Q04773)	1113.45	MDPELQAIR		21	N
42	5-methyltetrahydropteroyltriglutamate (P05694)	1114.61	VQSAVLGFPR	M	57	C
43	60S ribosomal protein L11-B (Q3E757)	1115.54	STKAQNPMR	M	32	N
44	40S ribosomal protein S10-B (P46784)	1115.55	MLMPKQER		49	-
45	40S ribosomal protein S15 (Q01855)	1126.61	SQAVNAKKR	M	57	N
46	Cell division control protein 11 (P32458)	1130.59	SGIIDASSALR	M	31	N
47	Ubiquitin-like modifier HUB1 (Q6Q546)	1134.58	MIEVVVNDR		26	C
48	U3 small nucleolar RNA-associated protein 14 (Q04500)	1141.65	AKKKSKSR	M	19	C
49	60S ribosomal protein L37-A (P51402)	1159.60	GKGTPSFGKR	M	39	C
50	Glycyl-tRNA synthetase 1 (P38088)	1170.62	SVEDIKKARAAVPFNR	M	28	N
51	40S ribosomal protein S11 (P26781)	1190.58	STELTVQSER	M	158	N
52	60S ribosomal protein L23 (P04451)	1205.58	SGNGAQGTKFR	M	109	N
53	Proteasome component C7-alpha (P21243)	1208.54	SGAAAASAAGYDR	M	32	N
54	40S ribosomal protein S2 (P25443)	1225.61	SAPEAQQQKR	M	108	N
55	DNA-directed RNA polymerase II (P16370)	1225.63	SEEGPQVKIR	M	25	N
56	Ubiquitin carboxyl-terminal hydrolase 6 (P43593)	1240.57	SGETFEFNIR	M	12	N
57	DNA-directed RNA polymerase II subunit RPB1(P04050)	1246.63	VGQQYSSAPLR	M	10	-
58	ATP phosphoribosyltransferase (P00498)	1254.60	MDLVNHLTDR		36	-
59	26S proteasome regulatory subunit RPN6 (Q12377)	1269.66	SLPGSKLEEAR	M	28	N
60	Gamma-glutamyl phosphate reductase (P54885)	1272.64	SSSQQIAKNAR	M	18	N
61	Acetyl-CoA acetyltransferase (P41338)	1278.66	SQNVYIVSTAR	M	36	N
62	60S ribosomal protein L26-A (P53221)	1288.63	AKQSLDVSSDR	M	49	C
63	60S ribosomal protein L37-A	1305.65	SYNWGAKAKR	M	69	N
64	DNA-directed RNA polymerase I subunit RPA2 (P22138)	1305.74	SKVIKPPGQAR	M	45	N
65	Vacuolar amino acid transporter 3 (P36062)	1307.58	MNGKEVSSGSGR		24	-
66	Cystathionine beta-synthase (P32582)	1316.58	TKSEQQADSR	M	14	-
67	Phosphate system positive regulatory protein PHO81 (P17442)	1325.68	MKFGKYLEAR		9	-
68	UPF0303 protein YBR137W (P38276)	1337.79	VVLDKKLLER	M	25	C
69	Reduced viability upon starvation protein 167 (P39743)	1352.71	SFKGFTKAVSR	M	23	N
70	Hexokinase-2 (P04807)	1355.77	VHLGPKKPQAR	M	58	C

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score	N^{α}-acetylation
71	40S ribosomal protein S4 (P05753)	1357.79	ARGPKKHLKR	M	69	C
72	Protein PUF6 (Q04373)	1380.77	APLTKKTNGKR	M	12	C
73	ATP-dependent RNA helicase DBP3 (P20447)	1384.72	TKEEIADKKR	M	10	C
74	Transcription factor tau 138 kDa subunit (P34111)	1386.00	SSVGIASASLWLR	M	16	N
75	Phospho-2-dehydro-3-deoxyheptonate aldolase (P14843)	1386.64	MFIKNDHAGDR		36	C
76	Glucokinase GLK1 (P17709)	1401.65	SFDDLHKATER	M	28	N
77	Multiprotein-bridging factor 1(O14467)	1405.65	SDWDTNTIIGSR	M	9	N
78	Bud emergence protein 1 (P29366)	1405.78	MLKNFKLSKR		32	-
79	40S ribosomal protein S29-B (P41058)	1420.66	AHENVWFSHPR	M	158	C
80	60S ribosomal protein L20 (P0C2I0)	1430.73	AHFKEYQVIGR	M	269	C
81	60S ribosomal protein L18 (P07279)	1431.72	GIDHTSKQHKR	M	78	C
82	Uncharacterized protein YOR051C (Q08421)	1435.83	AKRPLGLGKQSR	M	25	C
83	AdoMet-dependent rRNA methyltransferase SPB1 (P25582)	1440.77	GKTQKKNSKGR	M	8	C
84	FK506-binding protein 1 (P20081)	1441.74	SEVIEGNVKIDR	M	19	N
85	Actin-binding protein (P15891)	1443.70	ALEPIDYTTHSR	M	58	N
86	Mannose-1-phosphate guanyltransferase (P15891)	1447.79	MKGLILVGGYGTR		36	C
87	60S ribosomal protein L15-A (P05748)	1452.72	GAYKYLEELQR	M	87	C
88	60S ribosomal protein L15-B (P54780)	1453.71	GAYKYLEELER	M	47	C
89	Protein LTV1 (P34078)	1463.74	SKKFSSKNSQR	M	25	N
90	F-actin-capping protein subunit beta (P13517)	1475.73	SDAQFDAALDLLR	M	30	N
91	Suppressor of kinetochore protein 1 (P52286)	1486.58	VTSNVVLVSGEGER	M	27	C
92	60S ribosomal protein L14-A (P36105)	1488.72	STDSIVKASNWR	M	47	N
93	Spore-specific protein YSW1 (P38280)	1490.72	SSLADTVEGSEAKR	M	9	N
94	Eukaryotic translation initiation factor 1A (P38912)	1509.83	GKKNTKGGKKGR	M	54	C
95	Protein BCP1 (Q06338)	1536.86	VQAIKLNDLKNR	M	12	C
96	Uroporphyrinogen decarboxylase (P32347)	1537.83	GNFPAPKNDLILR	M	10	C
97	Pyruvate decarboxylase isozyme 2 (P06169)	1538.80	SEITLGKYLFER	M	78	N
98	T-complex protein 1 subunit beta (P39076)	1548.74	SVQIFGDQVTEER	M	12	N
99	T-complex protein 1 subunit zeta (P39079)	1551.86	SLQLLNPKAESLR	M	10	N
100	Uncharacterized protein YNR034W-A (Q3E841)	1553.85	MKSSIPITEVLPR		12	C
101	Adenylosuccinate lyase (Q05911)	1572.71	PDYDNYTTPLSSR	M	45	C
102	Eukaryotic translation initiation factor 3 (P38249)	1575.85	APPPFRPENAIKR	M	78	C
103	Protein transport protein SEC23 (P15303)	1580.68	MDFETNEDINGVR		63	N
104	Deoxyuridine 5'-triphosphate nucleotidohydrolase (P33317)	1597.90	TATSDKVLKIQLR	M	12	C
105	Prefoldin subunit 3 (P48363)	1609.74	MDTLFNSTEKNAR		55	N
106	60S ribosomal protein L29 (P05747)	1617.80	AKSKNHTAHNQTR	M	87	C
107	Ribose-phosphate pyrophosphokinase 3 (P38689)	1643.90	PTNSIKLLAPDVHR	M	25	C
108	T-complex protein 1 subunit gamma (P39077)	1648.77	MQAPVVFMNASQER		14	N
109	60S ribosomal protein L27 (P0C2H7)	1653.13	AKFLKAGKVAVVVR	M	78	C
110	Anaphase spindle elongation protein (P50275)	1658.82	METATSSPLPIKSR		56	-
111	60S ribosomal protein L5 (P26321)	1670.79	AFQKDAKSSAYSSR	M	69	C
112	Serine/threonine-protein phosphatase PP1-2 (P32598)	1673.76	MDSQPVDVDNIIDR		32	N
113	Co-chaperone protein SBA1 (P28707)	1694.87	SDKVINPQVAWAQR	M	13	C
114	tRNA (guanine-N(7)-)-methyltransferase (Q12009)	1709.88	MKAKPLSQDPGSKR		21	C
115	Alanyl-tRNA synthetase, cytoplasmic (P40825)	1742.90	TIGDKQKWTATNVR	M	41	C
116	Eukaryotic translation initiation factor 3 (P40825)	1756.81	SEVAPEEIIENADGSR	M	71	N
117	40S ribosomal protein S27-A (P38711)	1760.94	VLVQDLLHPTAASEAR	M	45	C
118	U3 small nucleolar RNA-associated protein 11 (P34247)	1766.03	AKLVHDVQKKQHR	M	32	C
119	Histone H2A.1 (P04912)	1771.88	SGGKGGKAGSAAKASQSR	M	12	N
120	60S ribosomal protein L13-A (Q12690)	1776.00	AISKNLPILKNHFR	M	54	C
121	40S ribosomal protein S18 (P35271)	1781.94	SLVVQEQGSFQHILR	M	78	N
122	Phosphoribosylformylglycinamidine synthase (P38972)	1788.94	TDYILPGPKALSQFR	M	41	-
123	Transketolase 1 (P23254)	1790.94	TQFTDIDKLAVSTIR	M	58	C
124	Aspartate-semialdehyde dehydrogenase (P23254)	1795.00	AGKKIAGVLGATGSVGQR	M	47	C
125	Cell division control protein 48 (P25694)	1802.88	GEEHKPLLDASGVDPR	M	18	C
126	Spermidine synthase (Q12074)	1810.90	AQEITHPTIVDGWFR	M	24	C
127	Peptidyl-prolyl cis-trans isomerase (P14832)	1821.85	SQVYFDVEADGQPIGR	M	95	N
128	Proliferating cell nuclear antigen (P15873)	1823.91	MLEAKFEEASLFKR		47	C
129	60S ribosomal protein L16-B (P26785)	1830.00	SQPVVVIDAKDHLLGR	M	69	N
130	Ribosomal RNA-processing protein 1 (P35178)	1837.86	METSNFVKQLSSNNR		98	-
131	Uncharacterized GTP-binding protein (P53295)	1840.96	GIIDKIKAIEEEMAR	M	22	-
132	60S ribosomal protein L16-A (P53295)	1844.02	SVEPVVVIDGKGHLVGR	M	150	N
133	26S proteasome regulatory subunit RPN10 (P38886)	1848.95	VLEATVLVIDNSEYSR	M	54	C
134	Ankyrin repeat-containing protein (P53066)	1862.87	MNTEGASLSEQLLDAAR		18	N
135	DNA repair protein REV1 (P12689)	1862.93	GLVDLLDSDLEYSINR	M	21	C
136	Glycerol-3-phosphate dehydrogenase (Q00055)	1864.95	SAAADRLNLTSGHLNAGR	M	204	N
137	40S ribosomal protein S21-A (Q3E754)	1873.92	MENDKGQLVELYVPR		147	N
138	U6 snRNA-associated Sm-like protein LSm3 (P57743)	1882.97	METPLDLLKLNLDER		18	N
139	Vacuolar ATP synthase subunit H (P41807)	1915.95	GATKILMDSTHFNEIR	M	14	C
140	Nonhistone chromosomal protein 6B (P11633)	1961.06	AATKEAKQPKEPKKR	M	25	C

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score	N^{α}-acetylation
141	GTP-binding protein RBG1 (P39729)	1968.00	STTVEKIKAIEDEMAR	M	22	N
142	Kelch repeat-containing protein 3 (Q08979)	1978.08	AKKNKKDKEAKKAR	M	14	N
143	Heat shock protein homolog SSE1 (P32589)	1986.02	STPFGLDLGNNNSVLAVAR	M	58	N
144	Ribonucleoside-diphosphate reductase small chain 2 (P49723)	2031.98	MEAHNQFLKTFQKER		14	-
145	Low specificity L-threonine aldolase (P37303)	2061.04	TEFELPPKYITAANDLR	M	44	C
146	Adenosylhomocysteinase (P39954)	2076.06	SAPAQNYKIADISLAAFGR	M	47	N
147	6,7-dimethyl-8-ribityllumazine synthase (P50861)	2098.11	AVKGLGKPDQVYDGSKIR	M	24	C
148	Phosphoglycerate kinase (P00560)	2099.11	SLSSKLSVQDLDLKDKR	M	158	N
149	40S ribosomal protein S28-B (P0C0X0)	2123.20	MDSKTPVTLAKVIKVLGR		89	-
150	Cytochrome c iso-1 (P00044)	2150.14	TEFKAGSAKKGATLFKTR	M	87	C
151	40S ribosomal protein S28-A (Q3E7X9)	2150.21	MDNKTPVTLAKVIKVLGR		107	-
152	D-lactate dehydrogenase (P39976)	2176.16	TAAHPVAQLTAEAYPKVKR	M	65	C
153	Seryl-tRNA synthetase, cytoplasmic (P07284)	2285.14	MLDINQFIEDKGGNPELIR		47	-
154	Ribosome biogenesis protein RLP7 (P40693)	2285.15	SSTQDSKAQTLNSNPEILLR	M	153	N
155	RuvB-like protein 1 (Q03940)	2298.15	VAISEVKENPGVNSSNSGAVTR	M	21	C
156	Aromatic amino acid aminotransferase 1 (P53090)	2303.06	TLPESKDFSYLFSDETNAR	M	15	C
157	RNA annealing protein YRA1 (Q12159)	2314.14	SANLDKSLDEIIGSNKAGSNR	M	18	N
158	Protein BMH1 (P29311)	2362.13	STSREDSVYLAKLAEQAER	M	31	N
159	Nascent polypeptide-associated complex subunit alpha (P38879)	2363.24	SAIPENANVTVLNKNEKKAR	M	24	N
160	60S ribosomal protein L6-A (Q02326)	2363.24	SAQKAPKWYPSEDVAALKKTR	M	69	N
161	Ribose-5-phosphate isomerase (Q12189)	2389.25	AAGVPKIDALESLGNPLEDAKR	M	47	C
162	60S ribosomal protein L32 (Q12189)	2436.40	ASLPHPKIVKKHTKKFKR	M	54	C
163	T-complex protein 1 subunit delta (P39078)	2467.67	SAKVPSNATFKNKEKPQEVR	M	31	N
164	Protein MET17 (P06106)	2468.12	PSHFDTVQLHAGQENPGDNAHR	M	52	C
165	60S ribosomal protein L8-A (P17076)	2492.34	APGKKVAPAPFGAKSTKSNKTR	M	87	N
166	Glutamine synthetase (P32288)	2501.30	AEASIEKTQILQKYLELDQR	M	78	N
167	tRNA pseudouridine synthase 1 (Q12211)	2524.03	SEENLRPAYDDQVNEDVYKR	M	36	N
168	60S ribosomal protein L6-B (P05739)	2539.27	TAQQAPKWYPSEDVAAPKKTR	M	98	C
169	Cytochrome c iso-2 (P00045)	2561.35	AKESTGFKPGSAKKGATLFKTR	M	35	C
170	Leu/Val/Ile amino-acid permease (P38084)	2641.21	MLSSEDFGSSGKKETSPDSISIR		12	-
171	Phosphoglucomutase-2 (P37012)	2703.34	SFQIETVPTKPYEDQKPGTSGLR	M	187	N
172	Nascent polypeptide-associated complex subunit (Q02642)	2717.47	PIDQEKLAKLQKLSANNKVGGTR	M	25	C
173	40S ribosomal protein S0-A (P32905)	2776.44	SLPATFDLTPEDAQLLLAANTHLGAR	M	158	N
174	Actin (P32905)	2833.25	DSEVAALVIDNGSGMCKAGFAGDDA	M	108	C
175	H/ACA ribonucleoprotein complex subunit 2 (P32495)	2843.32	GKDNKEHKESKESKTVDNYEAR	M	44	C
176	Acetyl-CoA carboxylase (Q00955)	2866.25	SEESLFESSPQKMEYEITNYSER	M	65	N
177	40S ribosomal protein S20 (Q3E7X9)	2926.47	SDFQKEKVEEQEQQQQQIIKIR	M	48	N
178	60S ribosomal protein L25 (P04456)	2928.68	APSAKATAAKKAVVKGTNGKKALKV	M	89	C
179	Vacuolar ATP synthase subunit B (P16140)	2939.58	VLSDKELFAINKKAVEQGFNVKPR	M	45	C
180	Elongation factor 3A (P16521)	2973.49	SDSQQSIKVLEELFQKLSVATADNR	M	58	N
181	GTP-binding nuclear protein GSP1/CNR1 (P32835)	3184.68	SAPAANGEVPTFKLVLVGDGGTGKT	M	22	N
182	NADPH-dependent 1-acyldihydroxyacetone phosphate reductase	3241.72	SELQSQPKKIAVVTGASGGIGYEVTK	M	44	N
183	60S ribosomal protein L7-A (P05737)	3432.81	AAEKILTPESQLKKSKAQQKTAEQVA	M	48	C
184	Histone H2A.Z (Q12692)	2094.05	SGKAHGGKGKSGAKDSGSLR	M	25	N
185	Phosphomannomutase (P07283)	3135.60	SIAEFAYKEKPETLVLFDVDGTLTPA	M	87	N

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score	N^{α}-acetylation
1	Acetoacetate metabolism regulatory protein (Q06065)	615.33	TAINR	M	21	-
2	Phage shock protein A (P0AFM7)	620.35	GIFSR		20	C
3	UPF0352 protein yejL (P0AD24)	641.35	PQISR	M	24	C
4	50S ribosomal protein L24 (P60625)	641.39	AAKIR			-
5	30S ribosomal protein S2 (P0A7V2)	705.35	ATVSMR	M	21	-
6	Outer membrane usher protein afaC mature (P53517)	744.38	AESGIAR	SP	39	N
7	Transcriptional repressor protein korC (Q52331)	744.38	SDVNIR		19	-
8	Hypothetical protein ydiZ (P64480)	758.39	ASGDLVR	M	23	C
9	Ribosome recycling factor (P0A805)	775.39	MISDIR		5	-
10	Hypothetical protein ydjI (P77704)	775.39	MLADIR	met ox	17	C
11	Carbon storage regulator (P69913)	787.46	MLILTR		33	C
12	Hypothetical protein yncE mature (P76116)	789.37	AEEMLR	M	23	C
13	Galactitol-specific phosphotransferase enzyme IIA component	790.43	TNLFVR	M	30	C
14	Methylglyoxal synthase (P0A732)	791.38	MELTTR		21	C
15	30S ribosomal protein S21 (P68681)	794.50	PVIKVR	M	23	C
16	Hypothetical protein yciW (P76035)	799.43	MLS-PIR		21	-
17	Hypothetical protein yebV	804.41	MKTSVR		20	C
18	Stringent starvation protein A (P0ACA3)	812.45	AVAANKR	M	39	C
19	Hypothetical protein ydcD (P31991)	814.47	MLQIIR		23	C
20	Lysine-arginine-ornithine-binding periplasmic protein mature (P09551)	826.45	ALPETVR	SP	19	C
21	Succinate dehydrogenase flavoprotein subunit (P0AC41)	826.47	MKLPVR		32	C
22	phosphoribosylformylglycinamidine synthase (Q8XA46)	833.41	MMEILR		22	-
23	Enoyl-[acyl-carrier-protein] reductase (P0AEK4)	847.45	GFLSGKR	M	35	C
24	Protein traP (P41068)	862.36	ANNMS-SR		10	C
25	Endonuclease III (P0AB84)	872.45	MNKAKR		22	C
26	GTP-binding protein typA/BipA (P32132)	872.48	MIEKLR		30	C
27	D-erythro-7,8-dihydroneopterin triphosphate epimerase (P0AC19)	880.51	AQPAAIIR	M	18	C
28	Aconitate hydratase 2 (P36683)	881.40	MLEEYR		25	C
29	DNA-invertase from lambdoid prophage e14 (P03014)	892.48	MLIGYVR		21	C
30	Dihydrodipicolinate reductase (Q8FLB4)	897.41	MHDANIR		35	-
31	Porphobilinogen deaminase (Q8XAP3)	901.47	MLDNVLR		23	C
32	Glutamyl-tRNA synthetase (Q8XBN2)	901.51	MKIKTR		24	-
33	Endonuclease V (P68739)	904.43	MDLASLR		29	-
34	Putative tagatose 6-phosphate kinase gatZ (P37191)	915.52	MKTLIAR		50	C
35	Cell division protein ftsY (P10121)	926.52	AKEKKR	M	25	C
36	Elongation factor G (P0A6M8)	926.53	ARTTPIAR	M	22	-
37	Chaperone protein sfmC mature (P77249)	927.51	AGGIALGATR	SP	19	C
38	Probable enoyl-CoA hydratase paaF (P76082)	928.49	S-ELIVS-R	M	46	-
39	Chaperone protein htpG (P0A6Z4)	932.44	MKGQETR		24	C
40	Transposase for insertion sequence element IS3411 (P11257)	937.56	PLLDKLR	M	35	C
41	30S ribosomal protein S12 (P0A7S5)	941.53	ATVNQLVR	M	9	C
42	Cell division protein ftsQ (P06136)	943.47	S-QAALNTR	M	5	-
43	Histidine biosynthesis bifunctional protein hisIE (Q8FG47)	946.03	MLTEQQR		46	-
44	Rod shape-determining protein mreB (P0A9X4)	947.53	MLKKFR		63	-
45	Protein tusB (Q8X885)	948.49	MLHTLHR		20	N
46	HTH-type transcriptional regulator dsdC (P46068)	953.53	EPLREIR	M	31	-
47	Succinate dehydrogenase hydrophobic membrane anchor protein	957.49	VS-NASALGR	M	27	C
48	Protein elaB (P0AEH7)	965.42	SNQFGDTR	M	33	C
49	Ribose operon repressor (P0ACQ0)	974.49	ATMKDVAR	M	57	C
50	Alanyl-tRNA synthetase (P00957)	974.50	SKSTAEIR	M	32	C
51	Large-conductance mechanosensitive channel (P0A742)	975.54	SIIKEFR	M	36	C
52	Phosphoheptose isomerase (P63224)	979.48	MYQDLIR		42	C
53	30S ribosomal protein S10 (P0A7R7)	986.51	MQNQRIR		56	C
54	Hypothetical protein yeeL (P76349)	991.55	MFLASLLR		22	C
55	P53517) Outer membrane usher protein afaC mature	992.43	MRDTS-SGR		15	-
56	Protein cysZ (Q8FFB9)	992.49	VSSFTSAPR	M	56	C
57	DNA polymerase III beta subunit (P0A990)	993.50	MKFTVER		32	C
58	Histidyl-tRNA synthetase (P60906)	996.57	AKNIQAIR	M	40	C
59	Ribonuclease III (P0A7Y1)	997.54	MNPIVINR		23	C
60	Ankyrin-repeat protein B (P76205)	999.53	SQNDIIIR	M	72	-
61	Cation efflux system protein cusA (P38054)	1001.54	MIEWIIR		29	C
62	Maltoporin mature (Q8X5W7)	1005.47	VDFHGYAR	SP	52	C
63	Penicillin-binding protein 6 mature (P08506)	1008.52	TQYSSLLR	SP	50	C
64	Seryl-tRNA synthetase (P0A8L1)	1012.54	MLDPNLLR		36	C
65	Transcriptional repressor pifC (P10030)	1015.55	MLSQLNLR		32	C
66	Transaldolase B (P0A870)	1016.55	TDKLTSLR	M	44	C
67	5,10-methylenetetrahydrofolate reductase (P0AEZ1)	1020.48	SFFHASQR	M	7	-
68	UPF0304 protein yfbU (P0A8W8)	1021.43	MEMTNAQR		60	C
69	Hypothetical protein yfcL (P64540)	1023.47	MIAEFESR		49	C
70	Hypothetical protein yhiD (P0AFV3)	1024.55	MTAEFIIR		25	-

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score	N^{α}-acetylation
71	Multidrug translocase mdfA (P0AEZ0)	1027.54	QNKLASGAR	M	41	-
72	Hypothetical protein yaiA (P0AAN5)	1038.55	PTKPPYPR	M	7	C
73	Nickel transport system permease protein nikC (P0AFB0)	1042.49	MNFFLSSR		27	C
74	Hypothetical protein ymgC (P75994)	1043.47	MNNSIPER		24	-
75	30S ribosomal protein S14 (P0AG61)	1044.54	AKQSMKAR	M	32	C
76	Protein iscX (P0C0L9)	1045.52	GLKWTDSR	M	47	C
77	Succinate dehydrogenase hydrophobic membrane anchor protein	1046.52	MVSNASALGR		135	C
78	Hypothetical UPF0042 protein yhbJ (P0A894)	1046.56	MVLMIVSGR		12	-
79	Hypothetical protein yqgC (P64570)	1048.49	GITSAGMQSR	M	54	C
80	Elongation factor Tu (EF-Tu) (P0A6N1)	1048.52	SKEKFER	M	103	N
81	Hypothetical UPF0263 protein yciU (P0A8L7)	1049.43	MDMDLNNR		49	C
82	Galactoside O-acetyltransferase (P07464)	1050.43	MNMPMTER		35	C
83	Inner membrane transport permease ybhS (P0AFQ4)	1055.54	S-NPILSWR	M	18	-
84	Phage shock protein E mature	1066.52	AEHWIDVR	SP	30	C
85	Protein yifE (P0ADN2)	1067.49	AESFTTTNR	M	53	-
86	50S ribosomal protein L22 (P61177)	1068.54	METIAKHR		44	C
87	UPF0306 protein yhbP (P67762)	1074.57	METLIAISR		18	C
88	Protein cyaY (P27838)	1076.43	MNDSEFHR		35	C
89	Hypothetical protein ydcY (P64455)	1080.56	SHLDEVIAR	M	36	C
90	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase	1081.69	AVTKLVLVR	M	46	C
91	Lipoyl synthase (P60717)	1084.56	S-KPIVMER	M	33	-
92	Hypothetical fimbrial-like protein ydeR mature (P77294)	1086.57	ADVTITVNGR	SP	60	N
93	Protein ygiN (P0ADU2)	1086.61	MLTVIAEIR		55	C
94	Uroporphyrinogen decarboxylase (P29680)	1089.51	MTELKNDR		23	C
95	4-hydroxy-3-methylbut-2-enyl diphosphate reductase (P62623)	1096.61	MQILLANPR		27	C
96	Transcription elongation factor greA (P0A6W5)	1101.57	MQAIPMTLR		15	-
97	Protein slyX (P0A8R4)	1103.53	MQDLSLEAR		14	-
98	Inner membrane protein yjgQ (P0ADC7)	1103.54	MQPFGVLDR		41	C
99	Bifunctional purine biosynthesis protein purH (Q8X611)	1111.60	MQQRRPVR		11	C
100	Esterase yeiG (P33018)	1115.47	MEMLEEHR		58	C
101	Histidinol-phosphate aminotransferase (Q9S5G6)	1117.60	STVTITDLAR	M	54	-
102	Cell division topological specificity factor (P0A734)	1122.61	ALLDFFLSR	M	57	C
103	Iron-binding protein iscA (P0AAC8)	1132.57	SITLSDSAAAR	M	71	-
104	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (P62620)	1135.56	MHNQAPIQR		61	C
105	Hypothetical protein yfeK mature (Q47702)	1137.58	KLTAHEEAR	SP	29	C
106	Hypothetical UPF0020/UPF0064 protein ycbY (P75864)	1138.54	MNSLFASTAR		18	-
107	Enolase (P0A6P9)	1138.71	SKIVKIIGR	M	113	C
108	Bifunctional aspartokinase/homoserine dehydrogenase II (P00562)	1140.63	SVIAQAGAKGR	M	45	C
109	Stringent starvation protein B (P0AFZ4)	1143.56	MDLSQLTPR		24	C
110	Alanine racemase, biosynthetic (Q8FB20)	1157.62	MQAATVLINR		13	-
111	Multidrug translocase mdfA (P0AEY8)	1158.58	MQNKLASGAR		25	-
112	Probable monothiol glutaredoxin ydhD (P0AC69)	1158.62	STTIEKIQR	M	39	C
113	Thymidine phosphorylase (P07650)	1161.62	MFLAQEIIR		60	-
114	Curlin genes transcriptional activatory protein (P24251)	1162.61	TLPSGHPKSR	M	22	C
115	Hypothetical protein yjaG (P32680)	1162.63	MLQNPIHLR		46	C
116	Outer membrane protein W mature (P0A915)	1164.50	HEAGEFFMR	SP	69	C
117	Ribonucleoside-diphosphate reductase 1 subunit alpha (P00452)	1168.66	NQNLLVTKR	M	20	C
118	Putative colanic biosynthesis UDP-glucose lipid carrier transferase	1169.65	TNLKKRER	M	10	-
119	Arginine N-succinyltransferase (P0AE37)	1171.62	MMVIRPVER		33	C
120	UPF0098 protein ybhB (P12994)	1172.62	MKLISNDLR		69	C
121	Hydrogenase-4 component G (P77329)	1175.49	MNVNS-S-S-NR		11	C
122	Elongation factor P (P0A6N4)	1177.50	ATYYSNDFR	M	23	C
123	50S ribosomal protein L27 (P0A7M0)	1179.60	AHKKAGGSTR	M	18	C
124	Long-chain fatty acid transport protein mature (P10384)	1180.50	AGFQLNEFSSSGLGR	SP	80	C
125	30S ribosomal protein S20 (P0A7U9)	1182.67	ANIKSAKKR	M	25	C
126	30S ribosomal protein S3 (P0A7V5)	1188.64	GQKVHPNGIR	M	53	C
127	2,5-diketo-D-gluconic acid reductase B (P30863)	1190.64	AIPAFGLGTFR	M	52	C
128	10 kDa chaperonin (groES protein) (P0A6G1)	1192.61	MNIRPLHDR		39	C
129	Dihydroorotase (P05020)	1195.69	TAPSQVLKIR	M	46	C
130	Putative 6-pyruvoyl tetrahydrobiopterin synthase (P65870)	1195.69	STTLFKDFTFEAAHR	M	54	C
131	Protein traB (P27188)	1196.63	MNKVQIGAPR		12	C
132	30S ribosomal protein S9 (P0A7X5)	1199.52	AENQYYGTGR	M	22	C
133	33 kDa chaperonin (HSP33) (P0A6Y5)	1202.56	MPQHDQLHR		35	-
134	Deoxyribose-phosphate aldolase (P0A6L0)	1204.61	MTDLKASSLR		46	C
135	UPF0082 protein yebC (P0A8A0)	1210.58	AGHSKWANTR	M	20	C
136	Hypothetical protein ydcY (P64455)	1211.60	MSHLDEVIAR		55	-
137	Hypothetical protein in sul1 3'region (P26837)	1214.52	MDSEEPPNVR			-
138	Threonyl-tRNA synthetase (P0A8M3)	1223.65	PVITLPDGSQR	M	21	C
139	Phosphoribosylamine--glycine ligase (Q8FB69)	1226.68	MKVLVIGNGGR		83	C
140	Oxygen-insensitive NAD(P)H nitroreductase (P38489)	1228.68	MDIISVALKR		57	C

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score	N^{α}-acetylation
141	Peptide transport periplasmic protein sapA mature	1230.60	APESPPHADIR	M	26	-
142	Hypothetical protein yfaQ mature (P76463)	1238.69	EETPLQLVLR	SP	42	-
143	Beta-glucuronidase (P05804)	1240.66	MLRPVETPTR		15	C
144	Glyceraldehyde-3-phosphate dehydrogenase A (P0A9B4)	1244.69	TIKVGINGFGR	M	123	C
145	Phosphoribosylaminoimidazole-succinocarboxamide synthase	1249.61	MQKQAELYR	M	40	C
146	Hypothetical UPF0069 protein yjeK (P39280)	1249.68	AHIVTLNTPSR	M	14	C
147	Biotin carboxylase (P24182)	1255.70	MLDKIVIANR		47	C
148	Exodeoxyribonuclease V beta chain (P08394)	1256.62	SDVAETLDPLR	M	57	-
149	GMP synthase (P04079)	1264.60	MTENIHKHR		17	C
150	Uridylate kinase (P0A7E9)	1272.68	ATNAKPVYKR	M	26	C
151	Peptide chain release factor 2 (P66024)	1274.61	MFEINPVNNR		42	C
152	Copper resistance protein D (Q47455)	1274.67	MNDLIMIVIR		40	N
153	Asparaginyl-tRNA synthetase (P0A8M0)	1280.71	SVVPVADVLQGR	M	66	C
154	50S ribosomal protein L30 (P0AG53)	1284.74	AKTIKITQTR	M	30	C
155	UDP-N-acetylmuramate--L-alanine ligase (P17952)	1285.68	MNTQQLAKLR		10	C
156	Multifunctional CCA protein (P06961)	1289.72	MKIYLVGGAVR		45	-
157	DNA replication protein dnaC (P0AEF0)	1290.61	MKNVGDLMQR		11	C
158	Aspartate-semialdehyde dehydrogenase (P0A9Q9)	1290.65	MKNVGFIGWR		39	C
159	3-phosphoshikimate 1-carboxyvinyltransferase (Q8FJB6)	1299.69	MESLTLQPIAR		62	-
160	D,D-heptose 1,7-bisphosphate phosphatase (Q8FKZ1)	1299.72	AKSVPAIFLDR	M	42	C
161	Glutamate-1-semialdehyde 2,1-aminomutase (P23893)	1308.63	SKSENLYSAAR	M	47	C
162	5-methyltetrahydropteroyltriglutamate--homocysteine	1309.71	TILNHTLGFPR	M	76	C
163	Protein yciN (P0AB61)	1314.62	MNKETQPIDR		40	-
164	Hypothetical ABC transporter ATP-binding protein yheS (P63389)	1318.70	MIVFS-S-LQIR		31	-
165	Cysteinyl-tRNA synthetase (P21888)	1319.73	MLKIFNTLTR		44	-
166	Hypothetical protein yqjD (P64581)	1320.64	SKEHTTEHLR	M	102	C
167	Aspartate–ammonia ligase (P00963)	1334.70	MKTAYIAKQR		49	-
168	Pantoate–beta-alanine ligase (P31663)	1352.81	MLIIETLPLLR		62	-
169	Putative HTH-type transcriptional regulator yeaT (P76250)	1353.71	MNNLPLLNDLR		31	C
170	DNA-binding protein H-NS (P0ACF8)	1353.76	SEALKILNNIR	M	43	C
171	Pyridoxine kinase (P40191)	1362.06	SSLLLFNDKSR	M	31	C
172	Phosphoribosylglycinamide formyltransferase 2 (P33221)	1381.80	TLLGTALRPAATR	M	31	C
173	Protein ydjA (P0ACY1)	1384.75	MDALELLINRR		39	-
174	Hypothetical oxidoreductase yqhD (Q46856)	1385.65	MNNFNLHTPTR		46	C
175	HIT-like protein ycfF (P0ACE7)	1389.75	AEETIFSKIIR	M	49	C
176	Hypothetical ABC transporter ATP-binding protein yheS (P63389)	1390.78	MIVFSSLQIRR		13	-
177	Inner membrane protein yagU (P0AAA1)	1403.65	MNIFEQTPPNR		35	C
178	Regulator of sigma D (P0AFX4)	1403.67	MLNQLDNLTER		25	C
179	Elongation factor Ts (EF-Ts) (P0A6P1)	1412.79	AEITASLVKELR	M	285	C
180	60 kDa chaperonin (Protein Cpn60) (groEL protein) (P0A6F7)	1416.70	AAKDVKFGNDAR	M	13	C
181	Cysteine desulfuration protein sufE (Q7ADI5)	1420.83	ALLPDKEKLLR	M	22	C
182	Hypothetical UPF0001 protein yggS (P67080)	1422.70	MNDIAHNLAQVR		71	C
183	Thiosulfate-binding protein mature (P16700)	1424.70	TELLNSSYDVSR	SP	55	C
184	Probable holo-[acyl-carrier-protein] synthase 2 (Q8XAN7)	1430.76	MRIGTDIVEIAR		19	-
185	Glutaminyl-tRNA synthetase (Q8FJW4)	1431.71	SEAEARPTNFIR	M	40	C
186	Glycerophosphoryl diester phosphodiesterase mature (P09394)	1435.73	ADSNEKIVIAHR	SP	76	-
187	Outer membrane protein tolC mature (P02930)	1436.67	ENLMQVYQQAR	SP	21	C
188	Hypothetical UPF0267 protein yqfB (P67603)	1437.67	MQPNDITFFQR		30	C
189	S-ribosylhomocysteine lyase (P45578)	1441.72	PLLDSFTVDHTR	M	205	C
190	Argininosuccinate synthase (Q8X9M0)	1445.84	TTILKHLPVGQR	M	48	C
191	Trigger factor (TF) (Q8FKA7)	1446.71	MQVSVETTQGLGR		54	-
192	Peptide deformylase (P0A6K3)	1446.78	SVLQVLHIPDER	M	51	-
193	Mannitol-1-phosphate 5-dehydrogenase (P09424)	1454.75	MKALHFGAGNIGR		78	-
194	RNA polymerase sigma-E factor (P0AGB7)	1457.74	SEQLTDQVLVER		25	-
195	Tyrosyl-tRNA synthetase (Q8FH88)	1469.77	ASSNLIKQLQER	M	35	C
196	Hypothetical protein ybjS (P75821)	1472.80	MKVLVTGATSGLGR		18	C
197	DNA-directed RNA polymerase alpha chain (P0A7Z4)	1475.74	MQGSVTEFLKPR		15	C
198	Arginyl-tRNA synthetase (P11875)	1484.80	MNIQALLSEKVR		48	C
199	Protease III mature (P05458)	1498.74	ETGWQPIQETIR	SP	17	C
200	RNA polymerase sigma-E factor (P0AGB8)	1499.75	S-EQLTDQVLVER	M	41	-
201	Protein yfgD (P76569)	1508.81	TKQVKIYHNPR	M	45	C
202	50S ribosomal protein L3 (P60440)	1514.83	MIGLVGKKVGMTR		64	C
203	Galactitol-1-phosphate 5-dehydrogenase (P0A9S3)	1516.76	MKSVVNDTDGIVR		50	-
204	Aerobactin siderophore biosynthesis protein iucC (Q47318)	1526.69	MNHKDWDLVNR		17	C
205	Thioredoxin 2 (P0AGG6)	1531.67	MNTVCTHCQAINR		30	-
206	Protein icc (P0AEW4)	1541.81	MESLLTLPLAGEAR		21	C
207	Glutamate synthase [NADPH] small chain (P09832)	1551.77	SQNVYQFIDLQR	M	39	C
208	Adenosine deaminase (P22333)	1566.81	MIDTTLPLTDIHR		20	-
209	Fumarate reductase iron-sulfur protein (P0AC47)	1570.84	AEMKNLKIEVVR	M	27	-
210	Acyl-CoA thioester hydrolase yciA (P0A8Z0)	1577.82	STTHNVPQGDLVLR	M	37	C

	Protein (Accession)	Mass (Da)	Sequence	Processing	Mowse score	N°-acetylation
211	Glutaminase 2 (P0A6W2)	1583.83	AVAMDNAILENILR	M	40	-
212	Protein-export membrane protein secF (P0AG93)	1585.75	AQEYTVEQLNHGR	M	40	-
213	50S ribosomal protein L21 (P0AG49)	1591.76	MYAVFQSGGKQHR		71	C
214	Acidic protein msyB (P25738)	1595.75	TMYATLEEAIDAAR	M	144	C
215	Glucokinase (P0A6V8)	1604.82	TKYALVGDVGGTNAR	M	82	C
216	D-alanine--D-alanine ligase B (P07862)	1613.86	TDKIAVLLGGTSAER	M	33	C
217	Periplasmic beta-glucosidase mature (P33363)	1622.77	DDLFGNHPLTPEAR	SP	46	C
218	Hypothetical protein yieF (P0AGE6)	1625.94	SEKLQVVTLLGSLR	M	30	-
219	Tryptophanase (P0A853)	1627.78	MENFKHLPEPFR		42	C
220	Probable exodeoxyribonuclease VII large subunit (Q8FF64)	1632.82	MLPSQSPAIFTVSR		28	C
221	Ribonuclease Z (P0A8V0)	1632.85	MELIFLGTSAGVPTR		20	C
222	Multidrug resistance protein mdtG (P25744)	1642.74	SPCENDTPINWKR	M	11	-
223	Regulatory protein rop (RNA one modulator) (P03051)	1646.81	MTKQEKTALNMAR		62	C
224	Regulator of nucleoside diphosphate kinase (P0AFW4)	1653.87	SRPTIIINDLDAER	M	40	C
225	6-phosphogluconate dehydrogenase (P00350)	1659.84	SKQQIGVVGMAVMGR	M	85	C
226	Erythronate-4-phosphate dehydrogenase (P05459)	1662.81	MKILVDENMPYAR		47	-
227	1-(5-phosphoribosyl) imidazole-4-carboxamide i (Q9S5G4)	1666.93	MIIPALDLIDGTVVR		91	C
228	Phosphoglycerate kinase (P0A799)	1671.89	SVIKMTDLDLAGKR	M	97	C
229	3-oxoacyl-[acyl-carrier-protein] reductase (P0AEK2)	1676.86	MNFEGKIALVTGASR		71	-
230	Hypothetical protein ycdB mature (P31545)	1681.83	QKTQS-APGTLSPDAR	SP	36	C
231	HTH-type transcriptional regulator chbR (P17410)	1682.84	MMQPVINAPEIATAR		30	C
232	HAM1 protein homolog (Q8XCU5)	1696.93	MQKVVLATGNAGKVR		47	C
233	50S ribosomal protein L13 (P0AA12)	1703.89	MKTFTAKPETVKR		56	-
234	Ornithine decarboxylase, constitutive (P21169)	1706.83	MKSMNIAASSELVSR		21	-
235	Pyruvate dehydrogenase E1 component (P0AFG9)	1715.81	SERFPNDVDPIETR	M	29	-
236	Phospholipase A1 mature (P0A921)	1732.87	QEATVKEVHDAPAVR	SP	69	C
237	FolD bifunctional protein (P24186)	1736.98	AAKIIDGKTIAQQVR	M	69	-
238	Serine transporter (P0AAD6)	1738.80	METTQTSTIASKDSR		94	-
239	Protein trbF (P15068)	1739.90	MRENKSNPEL KIR		16	C
240	Protein hfq (Host factor-I protein) (P0A6X3)	1740.92	AKGQSLQDPFLNALR	M	71	C
241	Methionine aminopeptidase (P0AE18)	1755.91	AISIKTPEDIEKMR	M	28	C
242	2,3,4,5-tetrahydropyridine N-succinyltransferase (P0A9D8)	1761.87	MQQLQNIIETAFER		109	C
243	Hypothetical protein yceD (G30K) (P0AB28)	1763.00	MQKVKLPLTLDPVR		31	-
244	Protein-export protein secB (P0AG86)	1766.79	SEQNNTEMTFQIQR	M	69	C
245	Hypothetical protein yjgD (P0AF90)	1769.82	ANPEQLEEQREETR	M	34	C
246	Acetylornithine deacetylase (P23908)	1773.95	MKNKLPPFIEIYR		48	-
247	Inner membrane protein ygaP (P55734)	1776.94	ALTTISPHDAQELIAR	M	71	C
248	Homoserine O-succinyltransferase (P07623)	1776.99	PIRVPDELPAVNFLR	M	59	C
249	Protein syd (P0A8U1)	1778.85	MDDLTAQALKDFTAR		25	C
250	30S ribosomal protein S15 (Q8X9M2)	1778.90	SLSTEATAKIVSEFGR	M	50	C
251	Acetyl-coenzyme A synthetase (P27550)	1783.92	SQIHKHTIPANIADR	M	40	C
252	NifU-like protein (P0ACD4)	1803.84	AYSEKVIDHYENPR	M	58	C
253	Chaperone protein hscA (P0A6Z1)	1828.98	ALLQISEPGLSAAPHQR	M	39	C
254	Peptide chain release factor 1 (P0A7I2)	1847.00	MKPSIVAKLEALHER		111	-
255	Periplasmic oligopeptide-binding protein mature (P23843)	1851.01	ADVPAGVTLAEKQTLVR	SP	92	C
256	ATP synthase alpha chain (P0ABB0)	1872.96	MQLNSTEISELIKQR		59	-
257	GTP cyclohydrolase I (P0A6T7)	1874.03	PSLSKEAALVHEALVAR	M	92	C
258	Glutamyl-tRNA reductase	1887.09	TLLALGINHKTAPVSLR	M	36	C
259	Hypothetical protein yaiS (P71311)	1888.96	MDKVLDSALLSSANKR		17	C
260	Protein sirA (P0A890)	1898.94	TDLFSSPDHTLDALGLR	M	75	C
261	Hypothetical protein ybcW (P64436)	1900.88	MNKEQSADDPSVDLIR		23	-
262	Transcriptional repressor mprA (Protein emrR) (P0ACR9)	1912.91	MDSSFTPIEQMLKFR		62	-
263	Protein yfiA (P0AD49)	1916.97	TMNITSKQMEITPAIR	M	62	-
264	50S ribosomal protein L14 (P0ADY5)	1918.89	MIQEQTMLNVADNSGAR		120	-
265	Blue copper oxidase cueO mature (P36649)	1920.03	AERPTLPIPDLLTTDAR	SP	39	C
266	Outer membrane protein assembly factor yaeT mature (P0A942)	1927.98	AEGFVVKDIHFEGLQR	M	110	C
267	FKBP-type peptidyl-prolyl cis-trans isomerase slyD (P0A9K9)	1945.07	MKVAKDLVVSLAYQVR		199	C
268	Putative HTH-type transcriptional regulator yfgA (P27434)	2000.89	MNTEATHDQNEALTTGAR		118	C
269	Putative HTH-type transcriptional regulator yfgA (P27434)	2000.89	MNTEATHDQNEALTTGAR		140	C
270	MTA/SAH nucleosidase (P0AF12)	2018.04	MKIGIIGAMEEEVTLLR		20	-
271	Glycerol kinase (P0A6F4)	2022.01	TEKKYIVALDQGTTSSR	M	83	C
272	Isoleucyl-tRNA synthetase (P00956)	2038.97	SDYKSTLNLPETGFPMR	M	68	C
273	3-oxoacyl-[acyl-carrier-protein] synthase III (P0A6R0)	2039.04	MYTKIIGTGSYLPEQVR		59	-
274	Glucose-1-phosphatase mature	2049.00	QTVPEGYQLQQVLMMSR	SP	60	C
275	Glucose-1-phosphate thymidylyltransferase 2 (P61887)	2061.15	MKGIILAGGSGTRLHPITR		25	-
276	Peroxiredoxin osmC (P0C0L2)	2071.06	TIHKKGQAHWEGDIKR	M	39	C
277	Peptide methionine sulfoxide reductase msrA (P0A744)	2076.10	SLFDKKHLVSPADALPGR	M	65	C
278	Hypothetical UPF0289 protein yacF (P36680)	2084.02	MQTQVLFEHPLNEKMR		91	-
279	Carbamoyl-phosphate synthase small chain (P0A6F1)	2098.09	MIKSALLVLEDGTQFHGR		136	-
280	Nitrate/nitrite response regulator protein narL (P0AF28)	2104.06	SNQEPATILLIDDHPMLR	M	82	C

	Protein (Accession)	Mass (Da)	Sequence	Processing	N°-acetylation	Mowse score
281	Hypoxanthine phosphoribosyltransferase (P0A9M2)	2107.08	MKHTVEVMIPEAEIKAR		27	-
282	Ferric uptake regulation protein (P0A9A9)	2107.16	TDNNTALKKAGLKVTLPR	M	31	-
283	Cysteine synthase A (P0ABK5)	2110.11	SKIFEDNSLTIGHTPLVR	M	64	C
284	Probable glucarate transporter (Q46916)	2132.02	MS-S-LSQAAS-SVEKRTNAR		10	-
285	Beta-lactamase (P62593)	2132.30	HPETLVKVKDAEDQLQR	SP	313	C
286	Protein ycil (P0AB55)	2168.08	MLYVIYAQDKADSLEKR		85	-
287	UTP–glucose-1-phosphate uridylyltransferase (P0AEP3)	2174.28	AAINTKVKKAVIPVAGLGTR	M	35	-
288	Protein ygaD (P0A6G3)	2187.08	TDSELMQLSEQVGQALKAR	M	62	-
289	Glycine cleavage system H protein (P0A6T9)	2211.10	SNVPAELKYSKEHEWLR	M	78	C
290	30S ribosomal protein S5 (P0A7W3)	2272.22	AHIEKQAGELQEKLIAVNR	M	20	C
291	Lysine-sensitive aspartokinase III (P08660)	2313.09	SEIVVSKFGGTSVADFDAMNR	M	29	C
292	Aspartate aminotransferase (P00509)	2316.18	MFENITAAPADPILGLADLFR		215	C
293	6-phosphofructokinase isozyme 1 (P0A797)	2325.22	MIKKIGVLTSGGDAPGMNAAIR	ox	40	-
294	6-phosphofructokinase (Q8FBD0)	2325.22	MIKKIGVLTSGGDAPGMNAAIR		59	-
295	Transcription elongation protein nusA (P0AFF6)	2336.24	MNKEILAVVEAVSNEKALPR		132	C
296	tRNA (guanine-N(7)-)-methyltransferase (P0A8I5)	2356.16	MKNDVISPEFDENGRPLRR		27	C
297	Acetylglutamate kinase (P0A6C8)	2409.28	MNPLIIKLGGVLLDSEEALER		62	C
298	UPF0010 protein yeaD (P39173)	2423.39	MIKKIFALPVIEQISPVLSR		50	C
299	Uracil phosphoribosyltransferase (P0A8F0)	2481.40	MKIVEVKHPLVKHKLGLMR		43	C
300	Integration host factor alpha-subunit (P0A6X7)	2483.26	ALTKAEMSEYLFDKLGLSKR	M	25	C
301	DNA-binding protein HU-beta (P0ACF4)	2496.30	MNKSQLIDKIAAGADISKAAAGR		58	C
302	Lysyl-tRNA synthetase (Q8XD57)	2534.25	SEQHAQGADAVVDLNNELKTRR	M	26	C
303	Riboflavin synthase alpha chain (P0AFU8)	2576.33	MFTGIVQGTAKLVSIDEKPNFR		42	-
304	Trimethylamine-N-oxide reductase 1 mature (P33225)	2577.32	AQAATDAVIS-KEGILTGSHWGAIR	SP	63	C
305	Hypothetical protein ychN (P0AB52)	2577.33	MQKIVIVANGAPYGSESLFNSLR		72	C
306	Phosphoribosylformylglycinamidine cyclo-ligase (Q8XAC5)	2591.27	TDKTSLSYKDAGVDIDAGNALVGR	M	75	C
307	ATP synthase beta chain (P0ABB4)	2591.40	ATGKIVQVIGAVVDVEFPQDAVPR	M	48	C
308	Mannose permease IID component (P69807)	2595.22	S-EMVDTTQTTTEKKLTQSDIR	M	26	C
309	Glutamate/aspartate periplasmic-binding protein (P37902)	2600.36	DDAAPAAGSTLDKIAKNGVIVVGHR	SP	36	C
310	UPF0078 membrane protein ygiH (P60782)	2605.37	SAIAPGMILIAYLCGSISSAILVCR	M	42	C
311	Guanosine-5'-triphosphate,3'-diphosphate pyrophosphatase (P25552)	2650.26	GSTSSLYAAIDLGSNSFHMLVVR	M	15	C
312	Tryptophanyl-tRNA synthetase (P00954)	2690.37	TKPIVFSGAQPSGELTIGNYMGALR	M	62	C
313	Hypothetical protein yrfD (P45753)	2701.51	AFKIWQIGLHLQQQEAVAVAIVR	M	18	C
314	Thymidylate synthase (P0A884)	2734.33	MKQYLELMQKVLDEGTQKNDR		31	-
315	Universal stress protein G (P39177)	2740.35	MYKTIIMPVDVFEMELSDKAVR		21	-
316	Oligopeptide transport ATP-binding protein oppD (P76027)	2776.50	SVIETATVPLAQQQADALLNVKDLR	M	39	C
317	30S ribosomal protein S1 (P0AG69)	2849.44	TESFAQLFEESLKEIETRPGSIVR	M	154	C
318	NH(3)-dependent NAD(+) synthetase (P18843)	2872.58	TLQQQIIKALGAKPQINAEEEIRR	M	120	C
319	Adenylosuccinate synthetase (P0A7D4)	2922.51	GNNVVVLGTQWGDEGKGKIVDLLTER	M	67	C
320	Phosphoserine aminotransferase (P23721)	2956.51	AQIFNFSSGPAMLPAEVLKQAQQELR	M	46	C
321	Leucyl-tRNA synthetase (P07813)	2983.41	MQEQYRPEEIESKVQLHWDEKR		64	-
322	Protein yhbO (P45470)	2999.40	SKKIAVLITDEFEDSEFTSPADEFR	M	75	C
323	Phosphoserine aminotransferase (Q8XEA7)	3000.54	AQIFNFSSGPAMLPVEVLKQAQQELR		45	C
324	3-mercaptopyruvate sulfurtransferase (P31142)	3039.46	STTWFVGADWLAEHIDDPEIQIIDAR	M	24	C
325	Phosphoenolpyruvate-protein phosphotransferase (P08839)	3110.69	MISGILASPGIAFGKALLLKEDEIVIDR		56	C
326	Protein sseB (P0AFZ1)	3124.55	SETKNELEDLLEKAATEPAHRPAFFR	M	60	C
327	50S ribosomal protein L10 (P0A7J5)	3177.69	ALNLQDKQAIVAEVSEVAKGALSAVVADSR	M	49	C
328	Glycine betaine-binding periplasmic protein mature (P0AFM2)	3197.67	ADLPGKGITVNPVQSTITEETFQTLLVSR	SP	20	C
329	Transcription termination factor rho (P0AG30)	3199.61	MNLTELKNTPVSELITLGENMGLENLAR		33	C
330	Ferrichrome-iron receptor mature	3202.61	AVEPKEDTITVTAAPAPQESAWGPAATIAAR	SP	41	C
331	Protein recA (Recombinase A) (P0A7G6)	3240.68	AIDENKQKALAAALGQIEKQFGKGSIMR	M	62	C
332	D-3-phosphoglycerate dehydrogenase (P0A9T0)	3329.86	AKVSLEKDKIKFLLVEGVHQKALESLR	M	72	C
333	Adenine phosphoribosyltransferase (P69503)	3389.79	TATAQQLEYLKNSIKSIQDYPKPGILFR	M	39	C

Supplementary Data C: non N-terminal identifications

- **Mouse liver**
- ***S. cerevisiae***
- ***E. coli***

To screen for internal peptides, MS/MS data from the N-terminal preparations of mouse liver, *S. cerevisiae* and *E. coli* were used to search the entire SwissProt database through MASCOT. Fixed modification: lysine acetylation; variable modifications: N-terminal acetylation, oxidation of methionine; protease: Arg-C; missed cleavages: 1; peptide tolerance: 1.5Da, MS/MS tolerance: 0.6Da, instrument: ESI-TRAP, peptide charge: 1+, 2+ and 3+.

Supplementary data C: mouse liver non N-terminal identifications

	Protein	Mass (Da)	Sequence	Mowse score
1	Glutathione S-transferase P 1	1791.78	EAAQMDMVNDGVEDLR	77
2	Glutathione S-transferase Mu 1	1562.70	KHHLDGETEEER	64
3		1389.66	ADIVENQVMDTR	70
4	Argininosuccinate lyase	1932.08	INVLPLGSGAIAGNPLGVDR	39
5		1058.54	NDQVVTDLR	62
6	Aspartate aminotransferase	1713.93	VGNLTVVGKESDSVLR	45
7		1153.65	LVLGDNSPAIR	18
8	Carbamoyl-phosphate synthase	798.42	DADPILR	33
9	Argininosuccinate synthase	2099.00	EGAKYVSHGATGKGNDQVR	45
10		928.48	LKEYHR	29
11	Estradiol 17 beta-dehydrogenase 5	1199.65	SKIADGTVKR	49
12	Catalase	2230.07	GPLLVQDVVFTDEMAHFDR	47
13	Betaine--homocysteine S-methyltransferase 1	1097.57	AIAEELAPER	42
14	Adenosylhomocysteinase	1113.60	ESLIDGIKR	33
15	Fructose-bisphosphate aldolase B	1428.73	IKVENTEENRR	32
16	1,4-alpha-glucan branching enzyme	1504.76	RQFNLTDDDLLR	25

Supplementary data C: S. cerevisiae non N-terminal identifications

	Protein	Mass (Da)	Sequence	Mowse score
1	Enolase 1	1140.6	TGQIKTGAPAR	63
2		1862.9	SGETEDTFIADLVVGLR	28
3	60S ribosomal protein L2	1179.7	VDKPLLKAGR	56
4		1391.7	GIVKQIVHDSGR	62
5	60S ribosomal protein L26-A	943.50	DDEVLVVR	51
6	Elongation factor 2	1255.6	AGIISAAKAGEAR	67
7	Pyruvate decarboxylase isozyme 1	2038.9	WAGNANELNAAYAADGYAR	56
8	60S ribosomal protein L13-B	2994.4	DGKAPEAEQVLSAAATFPIAQPATDVEA	52
9	60S ribosomal protein L33-A	1415.7	NNLPAKTFGASVR	48
10	40S ribosomal protein S25-A	1768.8	AQHAVILDQEKYDR	48
11	Glutamine synthetase	1603.7	SVAKEGYGYFEDR	45
12	40S ribosomal protein S18	1273.6	AGELTQEELER	45
13	Suppressor protein STM1	2086.0	KGNNTANATNSANTVQKNR	42
14	40S ribosomal protein S3	1297.7	VTPTKTEVIIR	39
15	60S ribosomal protein L31	990.51	GVKGVEYR	36
16	40S ribosomal protein S30	2884.6	AGKVKSQTPKVEKTEKPKKPKGR	35
17	Histone H2A.1	1244.6	SAKAGLTFPVGR	33
18	60S ribosomal protein L5	1242.5	EGKTDYYQR	32
19	40S ribosomal protein S14-A	1414.6	IEDVTPVPSDSTR	25
20	Phosphoglycerate kinase	2014.0	VDFNVPLDGKKITSNQR	24
21	60S ribosomal protein L7-A	1172.6	GFGKINKQR	20

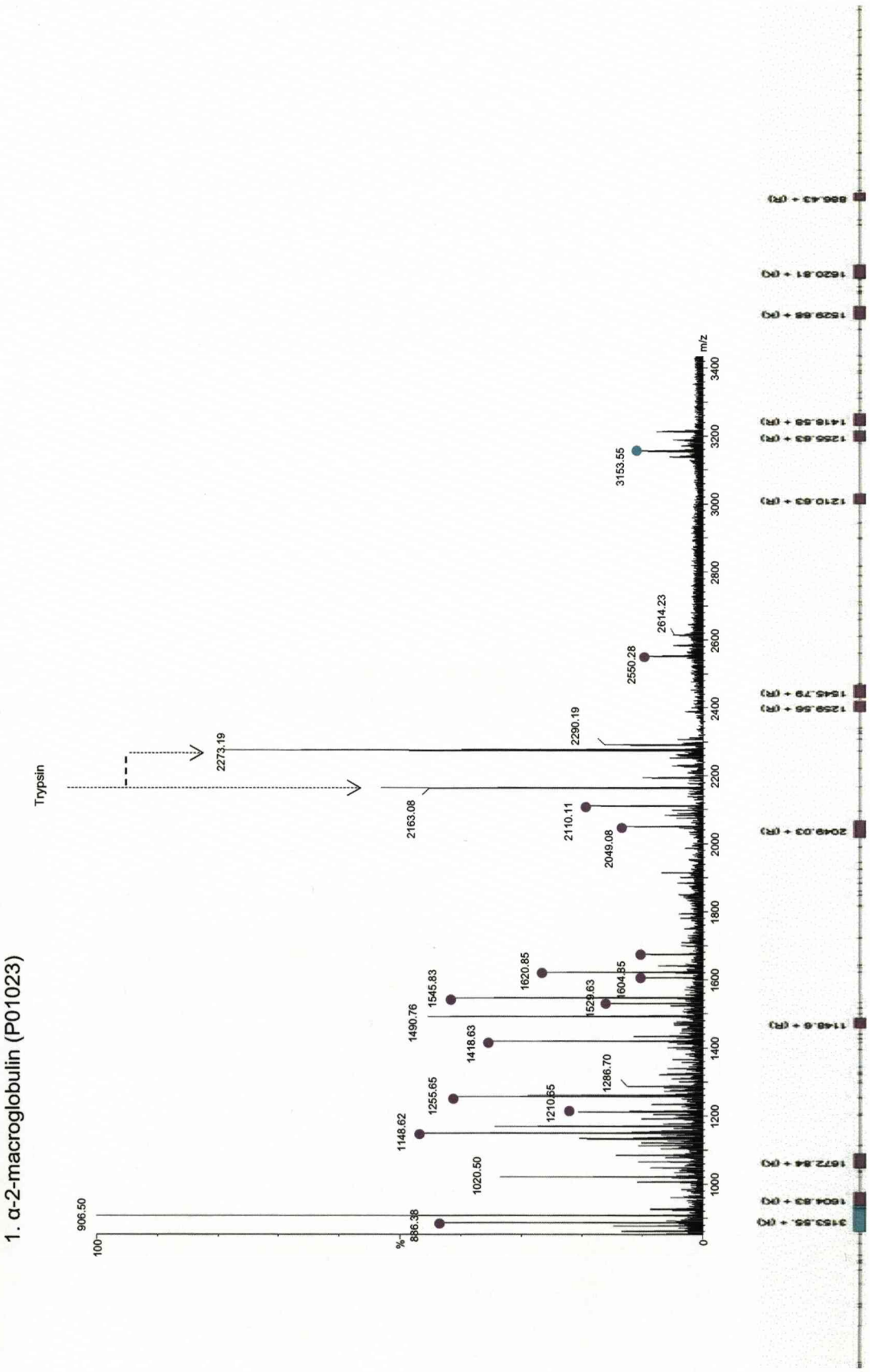
Supplementary data C: E. coli non N-terminal identifications

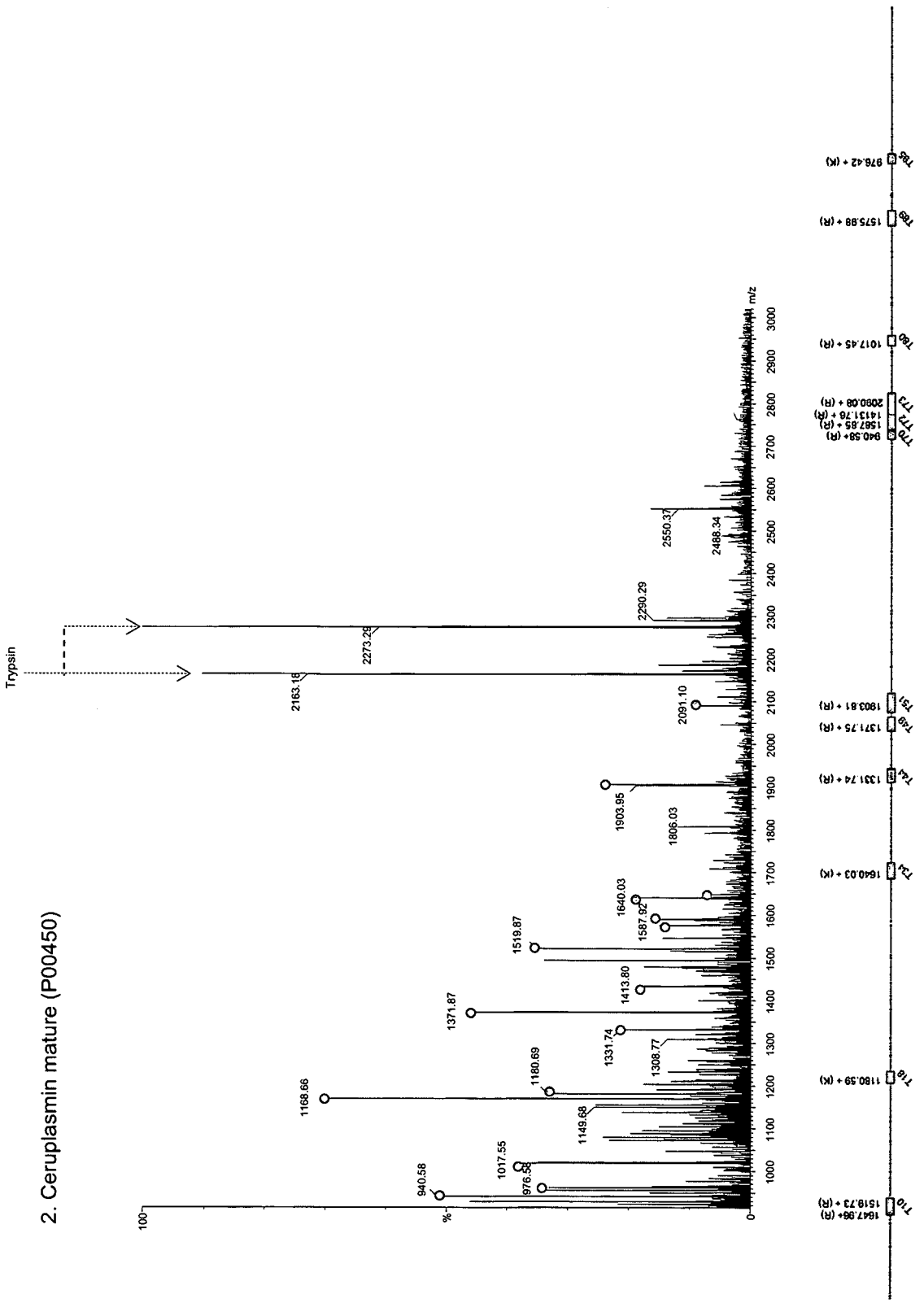
	Protein	Mass (Da)	Sequence	Mowse score
1	Chaperone clpB	1569.8	NKVTDAEIAEVLAR	90
2	Elongation factor Tu (EF-Tu)	1026.5	AGENVGVLLR	75
3	Beta-lactamase	1378.6	DTTMPAAMATTLR	56
4	"	1734.9	IVVIYTTGSQATMDER	46
5	"	1455.8	KLLTGELLTLASR	32
6	Chaperone protein dnaK	926.55	IAGLEVKR	52
7	"	2233.0	DAEANAEADRKFEELVQTR	50
8	Elongation factor G (EF-G)	2077.0	VYSGVVNSGDTVLNSVKAAR	75
9	"	1103.5	LAKEDPSFR	50
10	Phosphoglycerate kinase	2585.3	VKDYLDGVDVAEGELVVLENVR	27
11	"	1867.9	ADLNVPVKDGKVTSDAR	70
12	Transcription elongation protein nusA	1244.6	VQAVSTELGGER	60
13	Protein grpE	1286.6	VKAEMENLR	40
14	Phosphoribosylamine--glycine ligase	1226.6	FGDPETQPIMLR	82
15	Glyceraldehyde-3-phosphate dehydrogenase A	1494.8	VPTPNVSVVDLTVR	82
16	Hypothetical protein yqjD	1400.7	LGETGDAIAKQTR	74
17	Cysteinyl-tRNA synthetase	1319.7	ANENGESFVALVDR	62
18	Murein-lipoprotein	1152.5	LDNMATKYR	58
19	2,3-bisphosphoglycerate-dependent	1191.6	VIIAAHGNSLR	57
20	3-oxoacyl-[acyl-carrier-protein]	1126.6	VGLIAGSGGGSPR	50
21	Phosphoribosylformylglycinamidine synthase	1568.8	DVQTLKAKGDALAR	48
22	Peptide chain release factor 1	1666.8	AGTGGDEAALFAGDLFR	46
23	Phosphoribosylformylglycinamidine cyclo-ligase	1153.7	IKGVVKKTR	43
24	Pantoate--beta-alanine ligase	1058.5	AKDGLALSSR	30

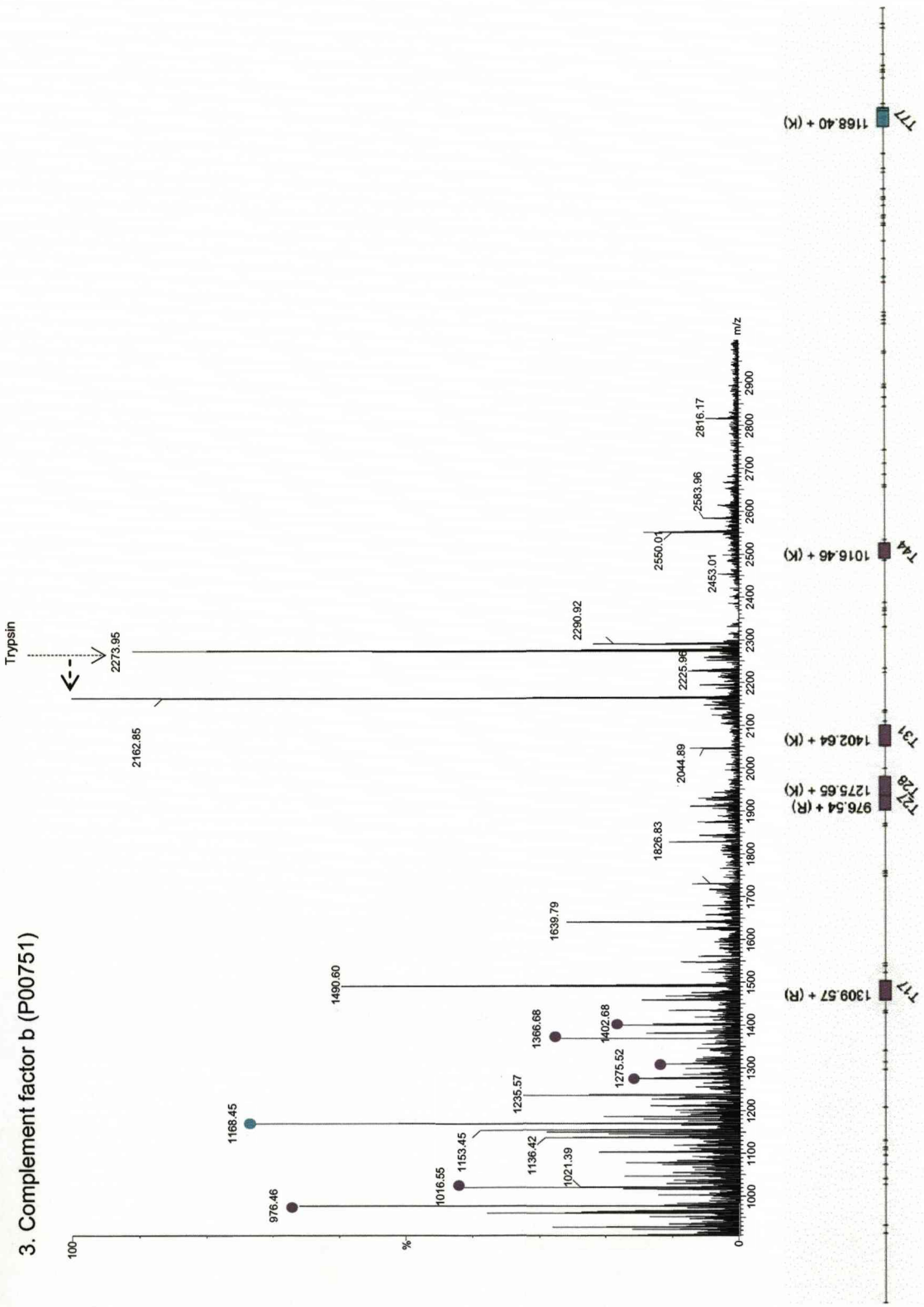
Supplementary Data D: Identification of human plasma proteins by PMF

- α-2-macroglobulin
- Ceruplasmin
- Complement factor b
- Serotransferin
- Serum albumin
- α -1-antitrypsin (P01009)
- Ig gamma-1 chain C region
- Apolipoprotein A-IV
- Ig kappa chain V-III region SIE
- Apolipoprotein A-I
- Transthyretin

Proteins from human plasma (15µg) were separated by 1-D SDS-PAGE and visualised with Coomassie. Gel plugs were excised from the dominant bands and subjected to in-gel proteolysis with trypsin (1:50 enzyme to substrate ratio). Peptide mixtures 1µl were spotted onto a MALDI target and allowed to air dry with 1µl of matrix solution. Samples were analysed by MALDI-ToF MS using a laser energy of 30%. The resulting peptide masses were imported into the MASCOT search engine. The taxonomy was restricted to *Homo sapiens;* fixed modification: carbamidomethylation of cysteine; variable modification: oxidation of methionine; protease: trypsin; missed cleavages: 1; peptide tolerance: 150ppm. Matched peptides are represented on the peptide coverage maps. Limit tryptic peptides are represented in purple and missed cleavages in green.

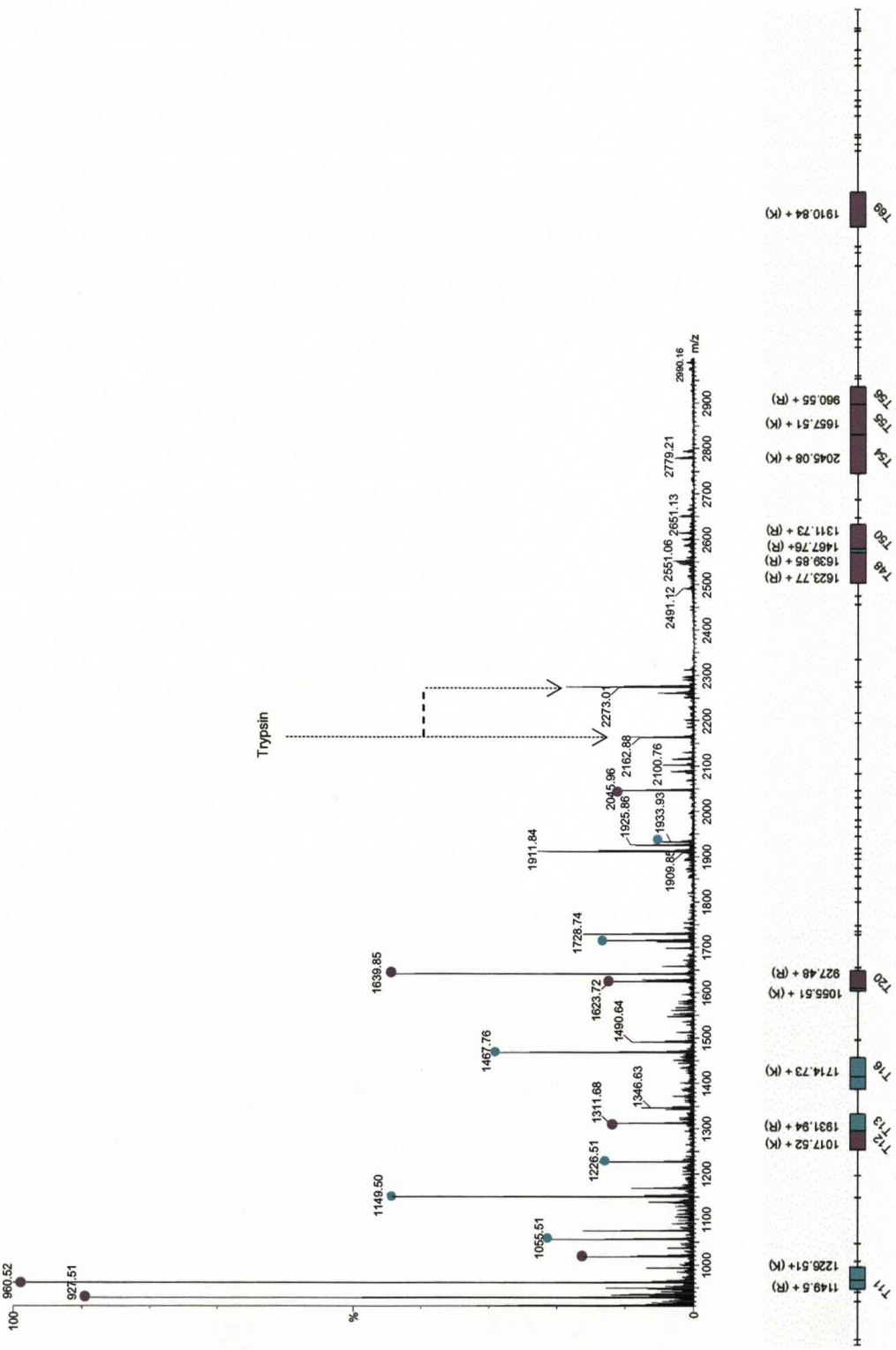

1. α-2-macroglobulin (P01023)
Trypsin
906.50
886.38
1148.62
1020.50
1255.65
1210.85
1418.63
1490.76
1545.83
1620.85
1604.85
1529.63
1286.70
2273.19
2163.08
2290.19
2110.11
2049.08
2550.28
2614.23
3153.55
m/z
100
%
0
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
3400
886.43 + (R)
1620.81 + (K)
1529.88 + (K)
1416.58 + (R)
1256.83 + (R)
1210.63 + (R)
1545.79 + (R)
1259.56 + (R)
2049.03 + (R)
1148.6 + (R)
1672.84 + (K)
1604.83 + (K)
3153.55 + (K)


2. Ceruplasmin mature (P00450)
Trypsin
100
%
0
m/z
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
2400
2500
2600
2700
2800
2900
3000
940.58
976.58
1017.55
1149.68
1168.66
1180.69
1308.77
1331.74
1371.87
1413.80
1519.87
1587.92
1640.03
1806.03
1903.95
2091.10
2163.18
2273.29
2290.29
2488.34
2550.37
T65 976.42 + (K)
T66 1575.88 + (R)
T69 1017.45 + (R)
T72 2080.08 + (R)
T73 1413.76 + (R)
T74 1587.85 + (R)
T75 940.58+ (R)
T76 1803.81 + (R)
T77 1371.75 + (R)
T78 1331.74 + (R)
T79 1640.03 + (K)
T80 1180.59 + (K)
T81 1518.73 + (R)
T82 1847.96+ (R)

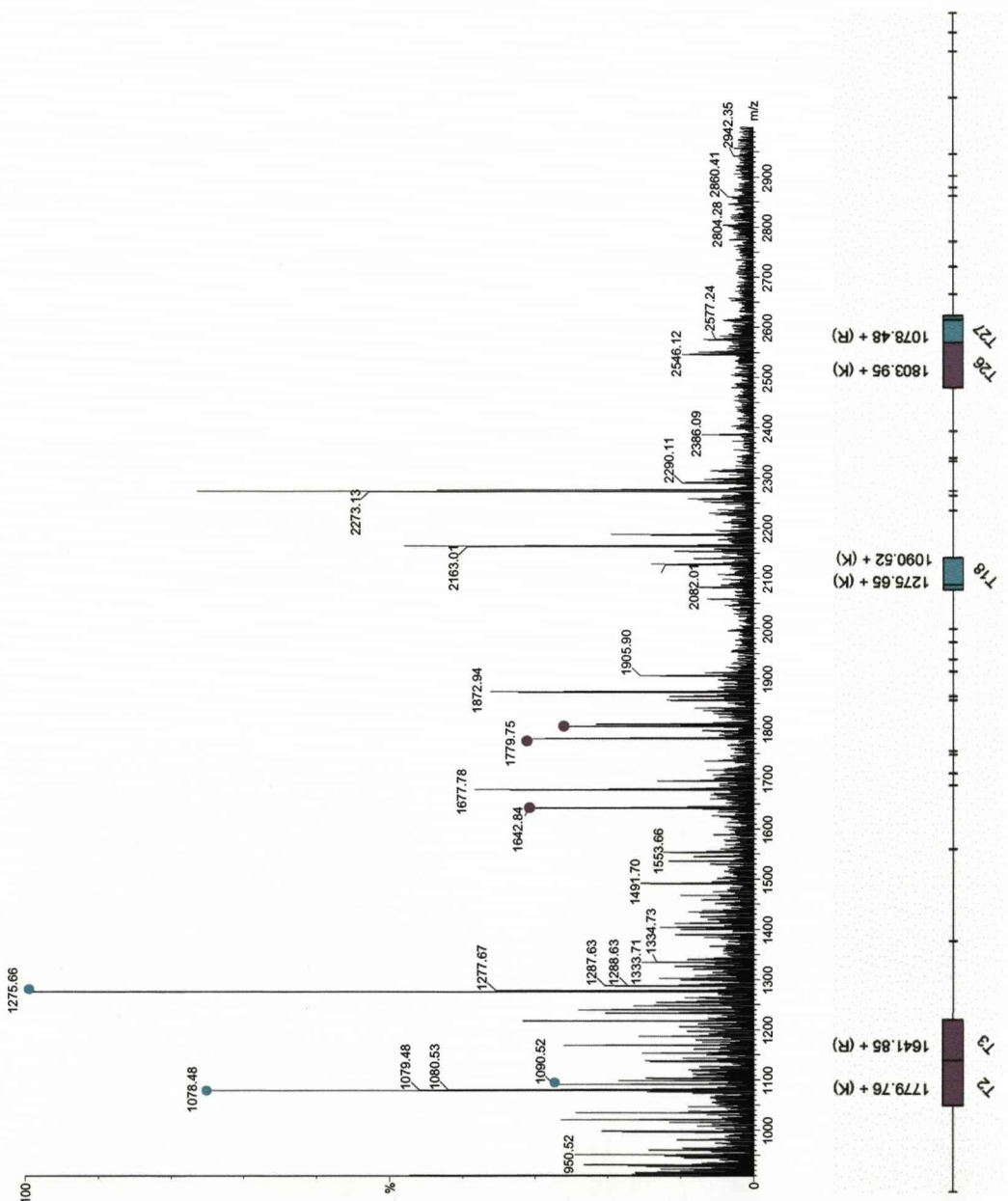

3. Complement factor b (P00751)
Trypsin
2273.95
2162.85
2290.92
2225.96
2550.01
2583.96
2816.17
2453.01
2044.89
1826.83
1639.79
1490.60
1366.68
1402.68
1275.52
1235.57
1168.45
1153.45
1136.42
1016.55
1021.39
976.46
100
%
0
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
2400
2500
2600
2700
2800
2900
m/z
T17
1168.40 + (K)
T14
1016.46 + (K)
T13
1402.64 + (K)
T12/28
1275.65 + (K)
976.54 + (R)
T17
1309.57 + (R)

4. Serotransferin (P02787)

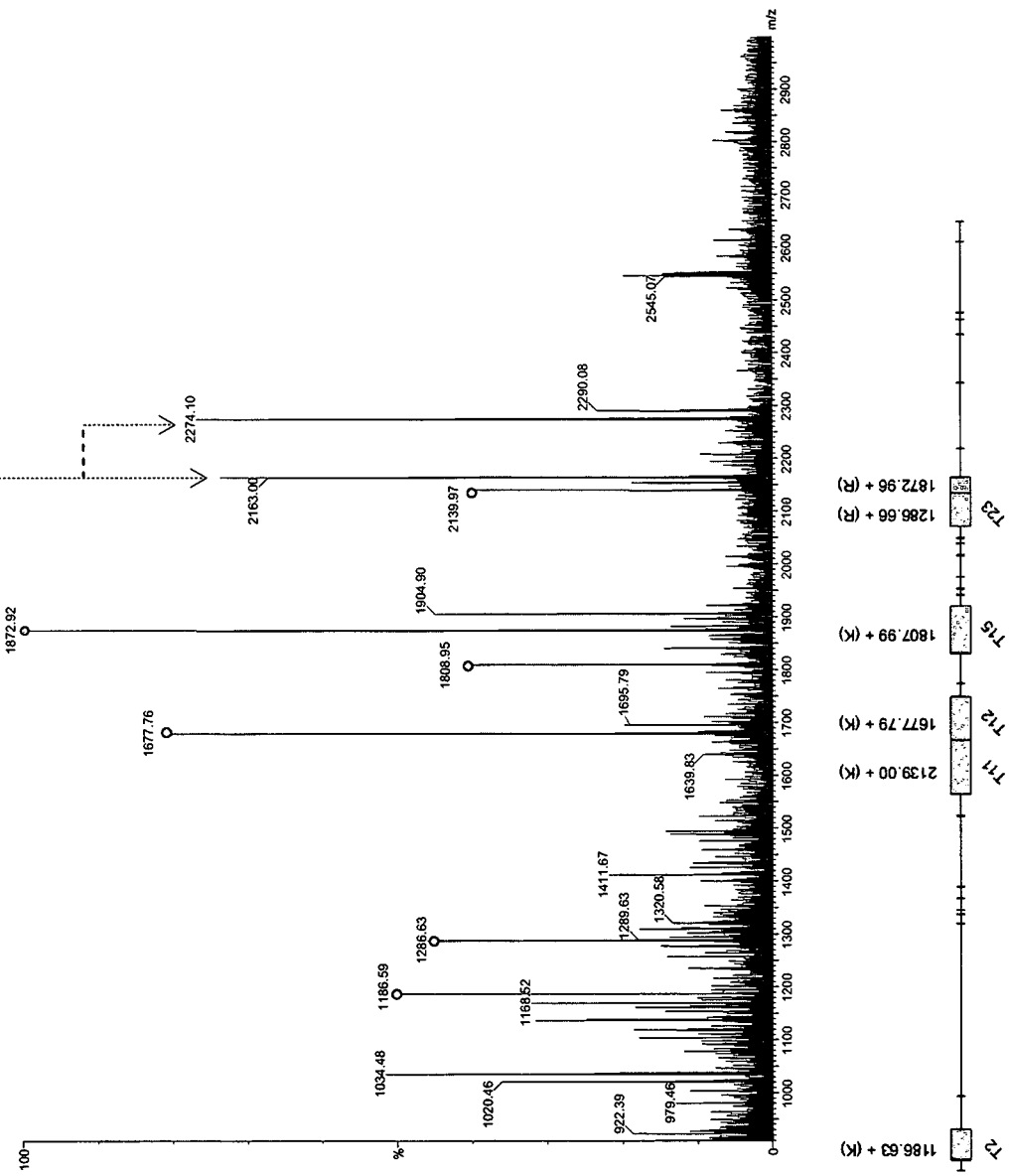
5. Serum albumin (P02768)



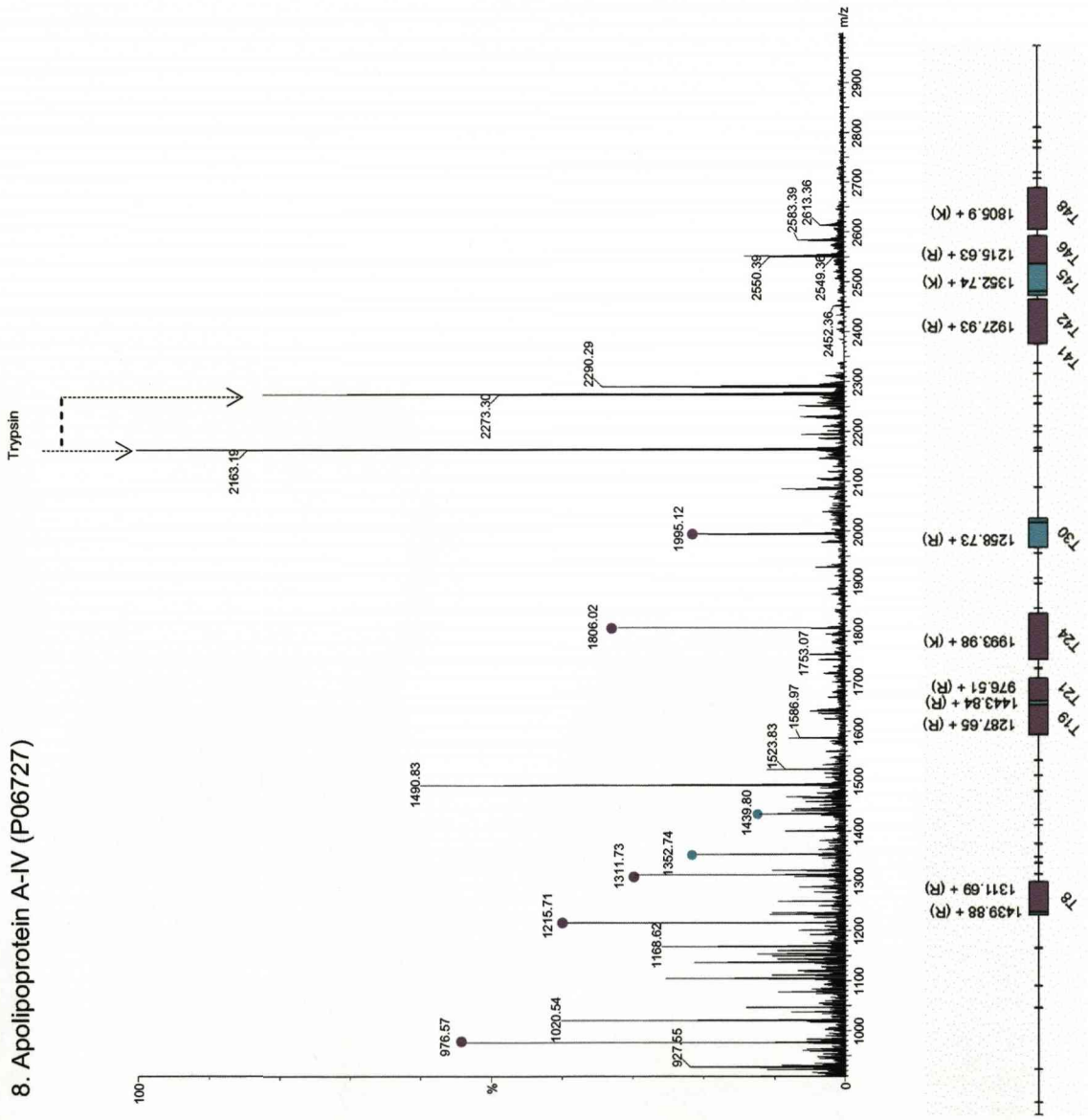
6. α -1-antitrypsin (P01009)

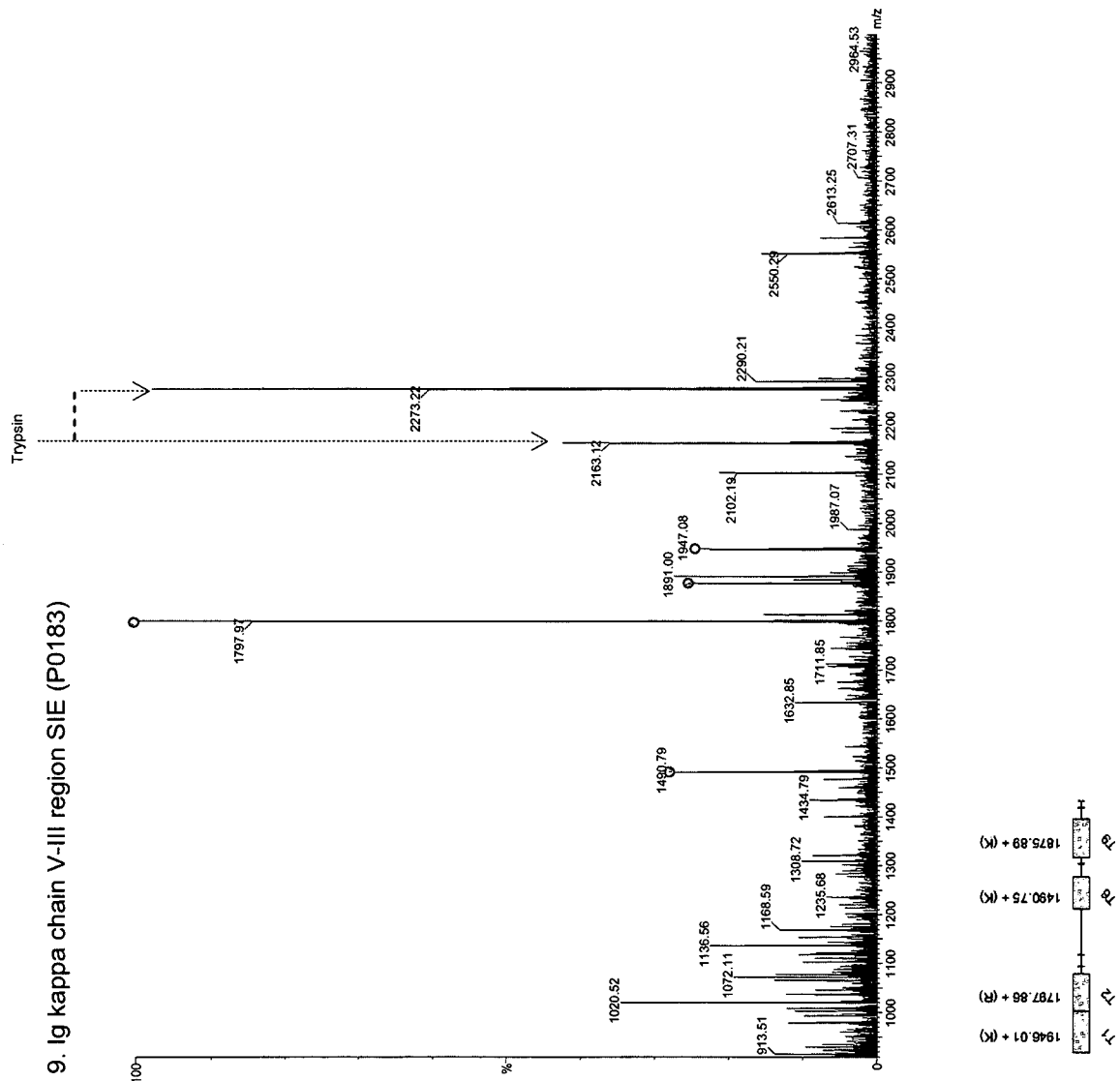


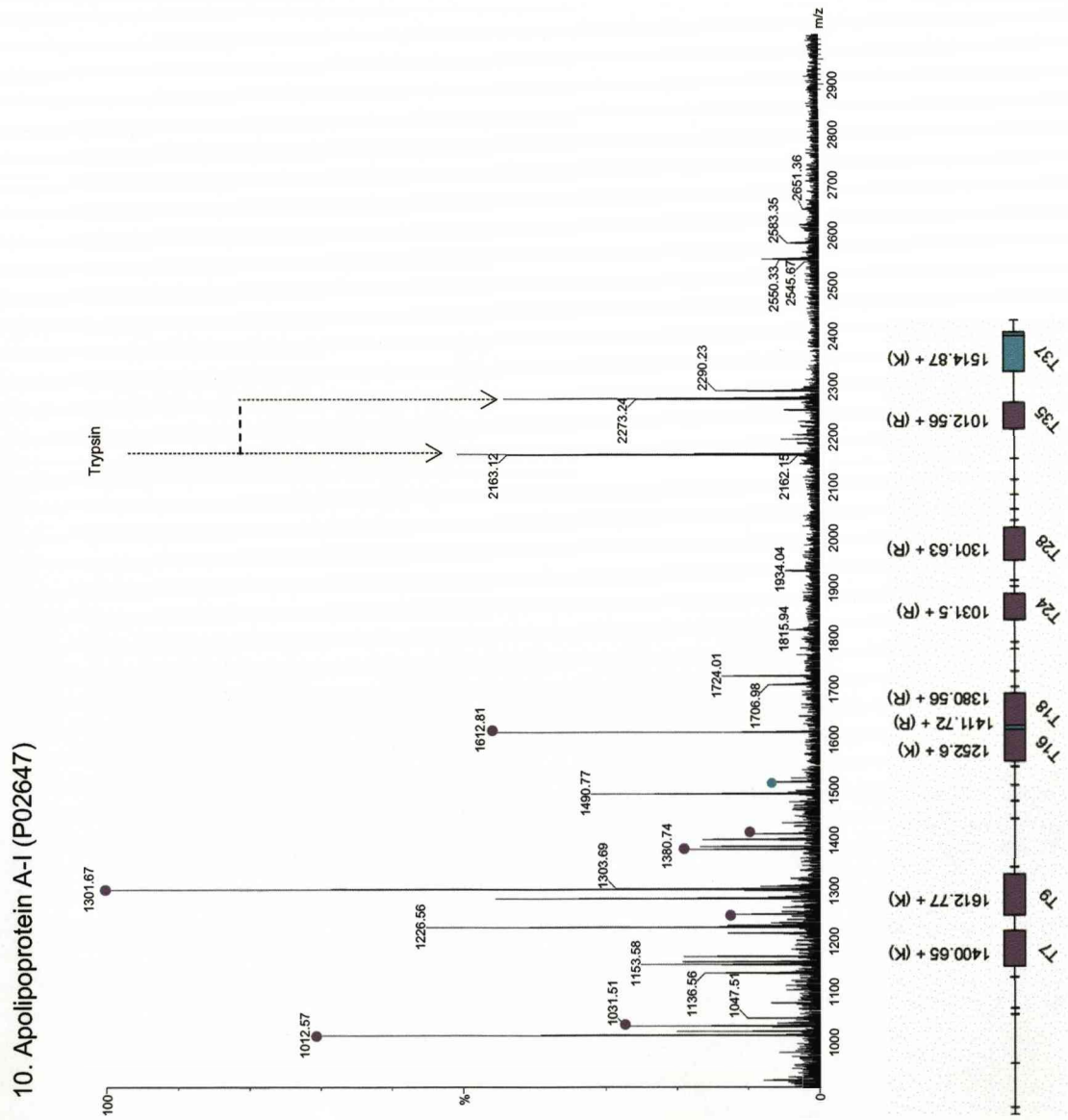
7. Ig gamma-1 chain C region (P01857)



8. Apolipoprotein A-IV (P06727)




9. Ig kappa chain V-III region SIE (P0183)
Trypsin
1946.01 + (K)
1787.88 + (R)
1490.75 + (K)
1875.89 + (K)
m/z
%
100
0
913.51
1020.52
1072.11
1136.56
1168.59
1235.68
1308.72
1434.79
1490.79
1632.85
1711.85
1797.97
1891.00
1947.08
2102.19
2163.12
2273.22
2290.21
2550.29
2613.25
2707.31
2964.53


10. Apolipoprotein A-I (P02647)
Trypsin
1301.67
1012.57
1226.56
1153.58
1031.51
1136.56
1047.51
1303.69
1490.77
1380.74
1612.81
1724.01
1706.98
1815.94
1934.04
2163.12
2162.15
2273.24
2290.23
2550.33
2545.67
2583.35
2651.36
m/z
%
100
0
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
2400
2500
2600
2700
2800
2900
T7
1400.65 + (K)
T9
1612.77 + (K)
T16
1252.6 + (K)
T18
1411.72 + (R)
1380.56 + (R)
T24
1031.5 + (R)
T28
1301.63 + (R)
T35
1012.56 + (R)
T37
1514.87 + (K)

Publications as a consequence of this thesis

McDonald, L., Robertson, D.H.L., Hurst, J.L. and Beynon, R.J. (2005) Positional proteomics: selective recovery and analysis of N-terminal peptides for identification proteomics. *Nature Methods* **2**, 955-957.

McDonald, L. and Beynon, R.J. (2006) Positional proteomics: preparation of amino-terminal peptides as a strategy for proteome simplification and characterisation. *Nature Protocols* **1**, 1790 - 1798.

Rivers, J., **McDonald. L.**, Edwards. I.J. and Beynon, R.J. (2008) 'Asparagine deamidation and the role of higher order protein structure. *Journal of Proteome Research* **7**, 921-927.

McDonald, L. and Beynon, R.J. *et al.* (2008) Mass isotope distribution analysis of amino acid residues (MIDAR): a novel acetylation reagent for proteomics. (in preparation for Molecular and Cellular Proteomics).

McDonald, L. and Beynon, R.J. *et al.* (2008) Utilisation of a positional proteomics strategy for biomarker identification in human plasma. (in preparation for Molecular and Cellular Proteomics).

McDonald, L. and Beynon, R.J. *et al.* (2008) Using positional proteomics to characterize protein N-terminal acetylation status. (in preparation for Journal of Biological Chemistry).

Siepen, J., **McDonald, L.**, Hubbard, S.J. and Beynon, R.J. (2008) Positional proteomics enhances proteome coverage: a benchmarking study. (In preparation for bioinformatics).

Contribution to each publication

- **McDonald, L.**, Robertson, D.H., Hurst, J.L. and Beynon, R.J. (2005) Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides. *Nat Methods* **2**, 955-957.

I was responsible for all experimental work, including method development, and preparation of the draft manuscript.

- **McDonald, L.** and Beynon, R.J. (2006) Positional proteomics: preparation of amino-terminal peptides as a strategy for proteome simplification and characterization. *Nat Protoc* **1**, 1790-1798.

I designed enhanced methodologies for N-terminal purification (NHS-Sepharose method) and demonstrated the use of the method in simplification of a complex proteome (*E. coli* cell lysate). I was responsible for preparation of the draft manuscript.

- Rivers, J., **McDonald, L.**, Edwards, I.J. and Beynon, R.J. (2008) Asparagine deamidation and the role of higher order protein structure. *J Proteome Res* **7**, 921-927.

Along with Ian Edwards, I made the initial observation of the atypical isotope distribution. I then went on to demonstrate the presence of the deamidated asparagine residue using methyl esterification.

- **McDonald, L.** and Beynon, R.J. *et al.* (2008) Mass isotope distribution analysis of amino acid residues (MIDAR): a novel acetylation reagent for proteomics. (In preparation for molecular and cellular proteomics).

I was responsible for all experimental work, including proof of principle data and the global N-terminal analysis, and preparation of the draft manuscript

- **McDonald, L.** and Beynon, R.J. *et al.* (2008) Utilisation of a positional proteomics strategy for biomarker identification in human plasma. (In preparation for molecular and cellular proteomics).

I was responsible for the initial N-terminal preparations of human plasma and data analysis, and for preparation of the draft manuscript

- **McDonald, L.** and Beynon, R.J. *et al.* (2008) Analysis of N-terminal acetylation status using positional proteomics. (In preparation for Journal of Biological Chemistry).

I conducted all of the experimental work and prepared the draft manuscript (work in progress).

- Siepen, J., **McDonald, L.**, Hubbard, S.J. and Beynon, R.J. (2008) Positional proteomics enhances proteome coverage: a benchmarking study. (In preparation for bioinformatics).

I conducted all of the experimental work and, along with Jennifer Siepen, I am presently involved in the preparation of the draft manuscript.

Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides

Lucy McDonald[1], Duncan H L Robertson[1], Jane L Hurst[2] & Robert J Beynon[1]

Bottom-up proteomics is the analysis of peptides derived from single proteins or protein mixtures, and because each protein generates tens of peptides, there is scope for controlled reduction in complexity. We report here a new strategy for selective isolation of the N-terminal peptides of a protein mixture, yielding positionally defined peptides. The method is tolerant of several fragmentation methods, and the databases that must be searched are substantially less complex.

Bottom-up proteomics operates at the level proteolytic peptides, generated from single proteins or from complex mixtures of proteins[1]. These peptides, generated by exhaustive proteolysis to limit peptides *in vitro*, are then analyzed by various mass spectrometric methods. Mass spectrometric analysis yields either the masses of a formally connected set of peptides that were all derived from a single protein (peptide mass fingerprinting) or, by tandem mass spectrometry, sequence-derived information that can identify the parent protein of a single peptide[2,3]. It can be argued that more peptides are analyzed than strictly necessary, and comprehensive proteomic analysis should focus on the minimal number of peptides that are required for protein identification. Methods such as ICAT[4] implicitly adopt this principle, in as much as the selective chemistry recovers only those peptides that contain at least one cysteine residue. Cysteine-mediated peptide recovery, however, is likely to abstract more than one peptide for each protein, and it is not possible to target the recovered peptide(s) positionally, as cysteine residues can occur anywhere in the protein sequence.

Positionally defined peptides would yield a substantial information gain in protein identification strategies. Most obviously, the two positional locations within every protein are the extreme ends—the N-terminal and the C-terminal peptides. Methods for recovery of C-terminal peptides have been reported, predominantly based on the ability of a catalytically disabled trypsin, anhydrotrypsin, to selectively bind peptides that terminate in a lysine or arginine residue[5,6]. There are several reports that indicate routes to selective recovery of N-terminal peptides, including specific N-terminal sequencing by mass spectrometry of gel-separated and blotted proteins[7], selective modification of N-terminal serine or threonine residues[8], modification of the hydrophobicity of a peptide mixture to preferentially expose N-terminal peptides by

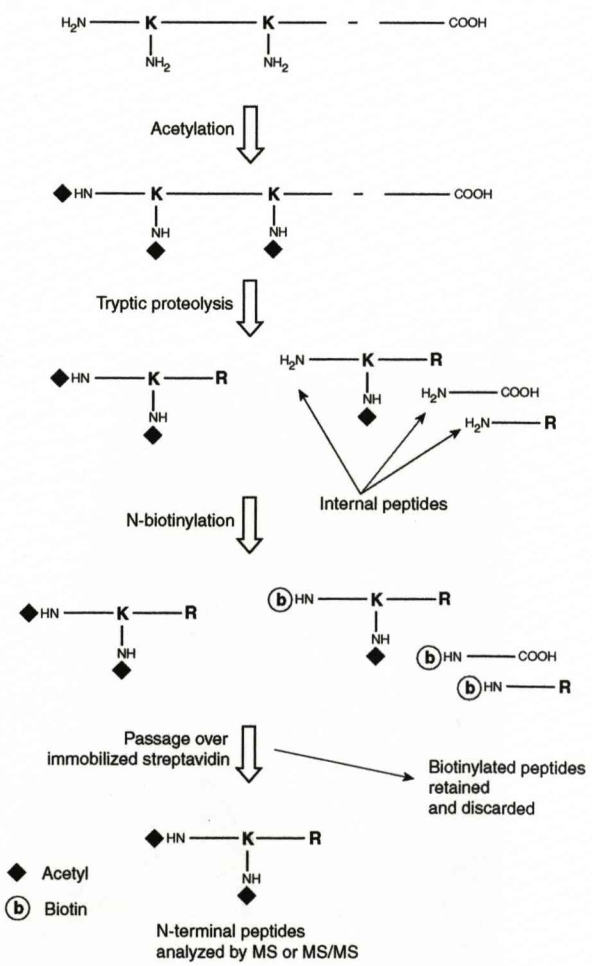


Figure 1 | Protocol for recovery of N-terminal peptides in a proteome. Free α- and ε-amino groups are acetylated before proteolysis (trypsin in the figures, but potentially any other fragmentation method), which is followed by biotinylation of proteolytically exposed α-amino groups. Subsequent subtractive binding to immobilized streptavidin creates a preparation enriched in those peptides that were originally derived from the N terminus, blocked by acetylation and therefore refractory to biotinylation.

[1]Protein Function Group and [2]Mammalian Behavior and Evolution Group, Faculty of Veterinary Science, University of Liverpool, Crown Street, Liverpool L69 7ZJ, UK. Correspondence should be addressed to R.J.B. (r.beynon@liv.ac.uk).

RECEIVED 23 AUGUST; ACCEPTED 3 OCTOBER; PUBLISHED ONLINE 18 NOVEMBER 2005; DOI:10.1038/NMETH811

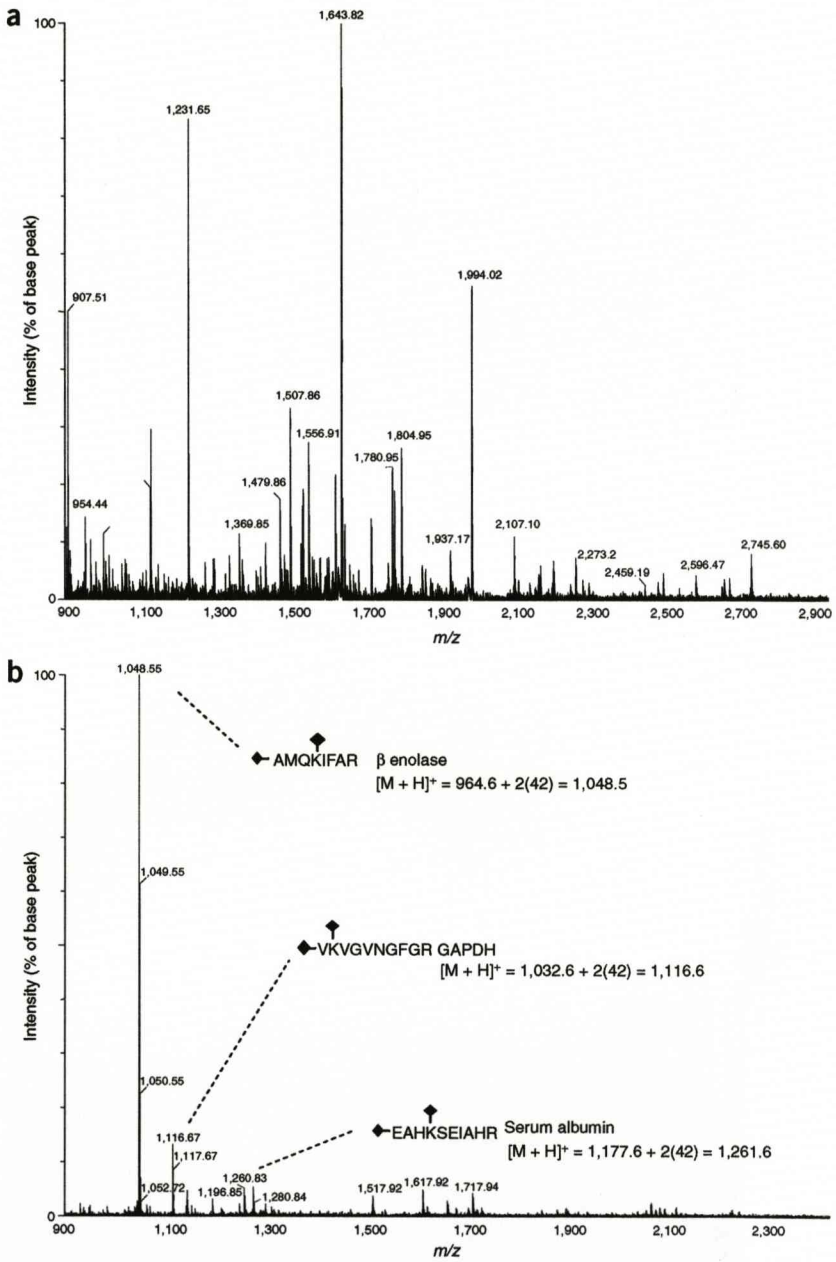


Figure 2 | Isolation of N-terminal peptides from soluble proteins of mouse skeletal muscle. (**a**,**b**) Skeletal muscle was homogenized in 10 ml of 20 mM sodium phosphate buffer (pH 8.0) and centrifuged for 45 min at 13,000*g*. The resultant supernatant fraction was used without further purification for preparation of N-terminal peptides (acetylation, tryptic proteolysis, N-biotinylation and subtractive purification). Detailed protocols are available in **Supplementary Methods** online. The entire tryptic digest of the mixture was analyzed by MALDI-ToF mass spectrometry (**a**). After application of the positional simplification protocol, the MALDI-ToF mass spectrum (**b**) contained major ions labeled in the figure, mass shifted by 42 Da through the addition of acetyl groups.

diagonal chromatography[9,10] and selective capture of all non–N-terminal peptides by amine scavenging beads[11,12]. We report here a new approach to selective recovery of the N-terminal-most peptides of a complex protein mixture. The method, based on subtractive removal of internal peptides, is not reliant on any particular endopeptidase cleavage—a flexibility that can compensate for the limitations in N-terminal peptide size distributions. Moreover, in contrast to other approaches[7,11,12], protein N termini that are naturally acetylated are automatically included in the analyte set. Indeed, if stable isotope–labeled acetic anhydride was used, it would be possible to identify and discriminate between naturally and chemically acetylated peptides. In brief, all available amino groups are blocked by acetylation. Subsequently, proteolysis generates new peptides, and all but the N-terminal peptide (whether naturally or artificially acetylated) expose a new amino group that is subsequently biotinylated. These biotinylated internal peptides are removed by recovery onto immobilized avidin or strepatividin, leaving behind the set of N-terminal peptides (**Fig. 1** and **Supplementary Methods** online, which contains a protocol for N-terminal peptide recovery and a description of the analysis of protein databases).

One of our major interests is in proteome dynamics in skeletal muscle[13–15]. The tryptic digest of the soluble protein fraction of mouse skeletal muscle contains peptides derived from a large number of proteins, and a matrix-assisted laser desorption/ionization–time-of-flight (MALDI-ToF) spectrum on an instrument of medium-level performance (resolution 12,000 FHWM (full width at half maximum)) yielded a detailed but complex mass spectrum (**Fig. 2a**). Owing to the complexity of the peptide mixture, we were unable to identify any N-terminal peptides in the spectrum. We passed the N-acetylated, trypsin-digested, biotinylated mixture over immobilized streptavidin. The unbound eluate gave a much simpler mass spectrum (**Fig. 2b**), and we were able to assign the highest intensity signals to true N-terminal peptides, confirmed by tandem mass spectrometry (**Supplementary Fig. 1** online). To test the method with a more complex mixture, we applied the same protocol to the soluble proteins of mouse liver. After purification of N-terminal peptides, the MALDI-ToF spectrum remained complex. By liquid chromatography–tandem mass spectrometry, and even without optimized separation or mass spectrometric analysis, many peptides (over 90) could immediately be assigned as N termini of mouse proteins (**Supplementary Fig. 2** and **Supplementary Table 1** online). As predicted, all terminated at C-terminal arginine residues. Moreover, the data were consistent with known or inferred N-terminal

processing (removal of initiator methionine, loss of signal peptide or propeptide) but in other cases have provided new information on N-terminal processing of liver proteins. All identifications were from a search of the entire database of mouse proteins rather than a restricted N-terminal database—there were virtually no peptides identified as internal sequences.

An analysis of extracted N-terminal peptides from mouse entries in Swissprot (**Supplementary Fig. 3** online) confirmed that over 85% of all proteins yielded an informative N-terminal peptide with trypsin digest, and that this value rose to almost 90% if we used two endopeptidases (trypsin and endopeptidase Glu-C). Thus, a substantial fraction of proteins in a proteome can be uniquely identified simply by the mass of the N-terminal peptide, using one or multiple endopeptidase digests. But the complexity of most N-terminal peptide preparations would require liquid chromatography–electrospray tandem mass spectrometry or liquid chromatography–MALDI tandem mass spectrometry. Even partial sequence data considerably enhance identification, and remove the need for multiple proteolytic digests—the residual 4% unidentifiable proteins represent sequences in the database that are either replicated entries or which represent paralogous proteins. There has recently been discussion about the uncertainty of 'one hit wonders' in proteomics, and we conjecture that part of the uncertainty relates to the lack of information about the location of the peptide in the parent proteins. A positional proteomics strategy anchors the peptides at a precise location within the parent protein, greatly reducing the search space for identification algorithms. As an average protein might be predicted to yield 50 tryptic peptides, the approximate reduction in search space is also 50-fold. Further, selective isolation and partial sequencing of N- and C-terminal peptides would allow virtually full length PCR-amplification of the cDNA corresponding to an expressed protein sequence.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

Supported by grants to R.J.B. & J.L.H. from the Natural Environment Research Council and the Biotechnology and Biological Sciences Research Council. We are grateful to M. Doherty for assistance with the mass spectrometry.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at
http://npg.nature.com/reprintsandpermissions/

1. Bogdanov, B. & Smith, R.D. *Mass Spectrom. Rev.* **24**, 168–200 (2004).
2. Standing, K.G. *Curr. Opin. Struct. Biol.* **13**, 595–601 (2003).
3. Steen, H. & Mann, M. *Nat. Rev. Mol. Cell Biol.* **5**, 699–711 (2004).
4. Smolka, M.B., Zhou, H., Purkayastha, S. & Aebersold, R. *Anal. Biochem.* **297**, 25–31 (2001).
5. Kasai, K. *J. Chromatogr.* **597**, 3–18 (1992).
6. Sechi, S. & Chait, B.T. *Anal. Chem.* **72**, 3374–3378 (2000).
7. Yamaguchi, M. *et al. Anal. Chem.* **77**, 645–651 (2005).
8. Chelius, D. & Shaler, T.A. *Bioconjug. Chem.* **14**, 205–211 (2003).
9. Gevaert, K. *et al. Nat. Biotechnol.* **21**, 566–569 (2003).
10. Martens, L. *et al. Proteomics* **5**, 3193–3204 (2005).
11. Kuhn, K. *et al. J. Proteome Res.* **2**, 598–609 (2003).
12. Kuhn, K. *et al. Proteomics* **5**, 2364–2368 (2005).
13. Doherty, M.K., Whitehead, C., McCormack, H., Gaskell, S.J. & Beynon, R.J. *Proteomics* **5**, 522–533 (2005).
14. Doherty, M.K. *et al. Proteomics* **4**, 2082–2093 (2004).
15. Hayter, J.R., Robertson, D.H., Gaskell, S.J. & Beynon, R.J. *Mol. Cell. Proteomics* **2**, 85–95 (2003).

Positional proteomics: preparation of amino-terminal peptides as a strategy for proteome simplification and characterization

Lucy McDonald & Robert J Beynon

Proteomics and Functional Genomics Group, Faculty of Veterinary Science, University of Liverpool, Crown Street, Liverpool L69 7ZJ, UK. Correspondence should be addressed to R.J.B. (r.beynon@liv.ac.uk).

Published online 16 November 2006; doi:10.1038/nprot.2006.317

We describe a protocol for selective extraction of the amino (N)-terminal-most peptide of a protein or a mixture of proteins after proteolysis. The first stage of the protocol blocks the free amino groups α and ε (the latter being lysyl residues) on the intact proteins by acetylation. In the second stage, proteolysis of the acetylated proteins yields a mixture of N-terminally acetylated (true N-terminal) and non-acetylated (internal and carboxy-terminal) peptides. Affinity capture of peptides bearing free amino groups using an immobilized amine-reactive reagent removes internal peptides from the mixture. The unbound fraction is highly enriched in N-terminal peptides, which can be analyzed without further treatment. This method is compatible with a range of proteolytic enzymes and fragmentation methods, and should take 2 d to complete. The N-terminal peptides can then be analyzed by mass spectrometry. This low cost, rapid method is readily adopted using off the shelf reagents.

INTRODUCTION

Although other strategies have been proposed, most proteomics studies are based on 'bottom-up' approaches and employ mass spectrometry (MS) for the analysis of limit proteolytic peptides (i.e., the products of proteolytic digestion of proteins in which all cleavable bonds have been hydrolyzed) that are derived, usually by tryptic hydrolysis, from single proteins or protein mixtures[1]. In all proteomics studies, there is a decision to be made about when to leave 'protein space' and to move into 'peptide space', with the transfer usually being a tryptic digestion (**Fig. 1a**). Most proteomics studies aim to simplify a complex proteome, to the level of subproteomes (e.g., a subcellular fraction) or individual proteins (e.g., proteins separated by 2D gel electrophoresis) before proteolysis. After separation of the proteins by, for example, 2D gel electrophoresis, it is reasonably inferred that the connectivity of the entire set of peptides is such that they are all derived from the same parent protein. However, any mixture of proteins generates a correspondingly mixed set of limit peptides, and the connectivity between different peptides therefore cannot be assumed. This would usually mean that tandem MS (MS/MS) is required, as further information must be extracted from a single peptide. The extreme resolution of MS and of the information obtained by MS/MS can be used to obviate exhaustive protein separation in favor of more global 'shotgun' approaches[2], in which a mixture of proteins (sometimes an entire proteome) is initially proteolyzed after which the complex mixture of peptides is separated before mass-spectrometric analysis, often with direct

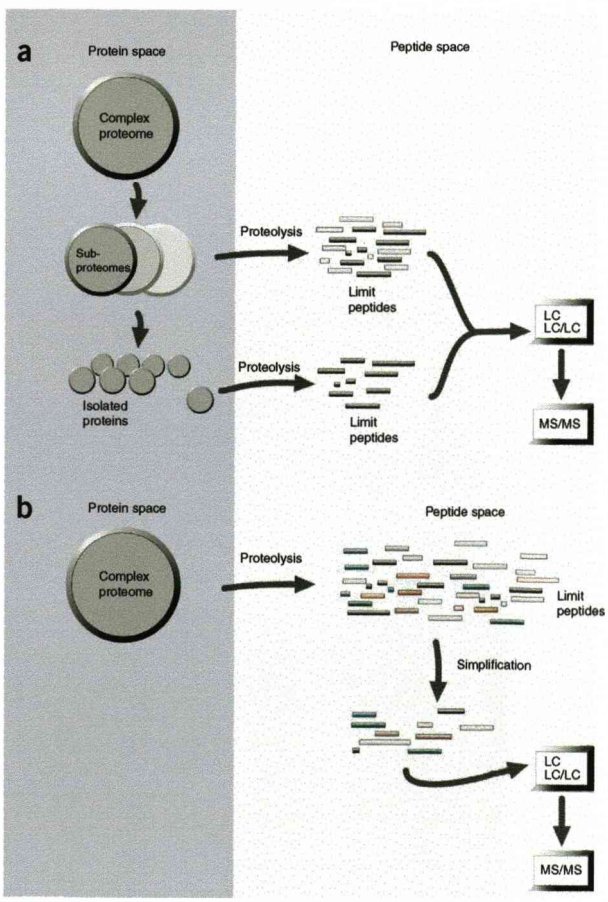


Figure 1 | Outline of a standard approach to protein identification. A complex proteome is simplified using a variety of separation techniques, either in protein space or, after proteolysis, in peptide space. The resulting subproteomes can be either proteolyzed directly or further subjected to separation into individual proteins before proteolysis. The resulting peptides are analyzed by mass spectrometry, with or without chromatographic separation, depending on the complexity of the final peptide mixture. (**a**) Strategy for selective/targeted peptide simplification. Proteolysis of a complete proteome creates a peptide mixture that is so complex that mass-spectrometric analysis is highly challenging. (**b**) By targeting specific structural regions on peptides, the mixture can be selectively purified in such a way that the majority of the proteome is discarded. LC, liquid chromatography.

coupling between the separation fluidics and the mass spectrometer. However, the complete repertoire of limit peptides generated from an entire proteome might contain many hundreds of thousands of peptides, which imposes a formidable analytical challenge. Even with the benefit of 2D chromatography (multidimensional protein identification technology or MUDPIT[3]), the mixture might deliver more peptides to the mass spectrometer than can feasibly be analyzed as the liquid stream flows through the electrospray source. Moreover, the data system is likely to direct MS analysis to more abundant proteins, therefore limiting dynamic range.

As a consequence of this complexity, there have been a number of attempts to simplify peptide mixtures. These approaches target specific chemical reactivities of the proteins or peptides within complex mixtures (**Fig. 1b**). The purpose of such methods is to eliminate the majority of the proteome-derived peptides but retain sufficient information necessary for analysis[4]. Lower abundance amino acids can be targeted as a means of extracting informative peptides. Methods such as isotope-coded affinity tagging (ICAT)[5] implicitly adopt this principle, using selective chemistry that recovers only those peptides that contain at least one cysteinyl residue and at the same time introducing tags for relative quantification. Immobilized metal-affinity chromatography (IMAC)[6], titanium dioxide-affinity chromatography[7], is used for the enrichment of phosphopeptides and lectin-affinity chromatography is used to target glycopeptides[8]. All of these simplification techniques have the potential to abstract more than one peptide for each protein and, in some instances, no peptides will be recovered. Furthermore, because the peptide could have been derived from any part of the parent protein sequence, no information on its location is obtained, which complicates the search strategy for identification proteomics. Finally, combined diagonal chromatography (COFRADIC) methods are able to enrich a subset of peptides according to their chemical composition (e.g., by the presence of methionyl residues) or enrich for amino (N)-terminal peptides[9–11].

An efficient strategy for proteome simplification would be the isolation of a single signature peptide from each protein in the proteome, optimizing the balance between analyte complexity and completeness of representation. As the disposition of protease cleavage sites within a protein is effectively random, the most obvious choice for a signature or proteotypic peptide would be the (amino N) or carboxy (C) terminus. Methods for recovery of C-terminal peptides have been reported, which are predominantly based on the capture of internal tryptic peptides using an anhydrotrypsin column[12,13]. Several strategies have been developed in order to selectively purify N-terminal peptides, including specific N-terminal sequencing by MS of gel-separated and blotted proteins[14], selective modification of N-terminal serine or threonine residues[15], and modification of the hydrophobicity of a peptide mixture to preferentially expose N-terminal peptides by diagonal chromatography[9].

We have previously described a strategy to selectively purify N-terminal peptides[16]. This approach required blocking of available amino groups on the intact proteins by acetylation, followed by proteolysis of the acetylated proteins. The resulting peptide mixture consisted of blocked N-terminal peptides mixed with internal peptides containing free amino groups on the proteolytically formed N-termini. Biotinylation of the peptide mixture using an N-hydroxysuccinimide (NHS) ester-derivative of biotin added a biotin moiety to the N-termini of the internal peptides, whereas true N-terminal peptides that had been previously blocked by acetylation would not be able to be biotinylated. The biotinylated peptides were removed by passing the mixture over streptavidin, and the unbound material containing the N-terminal peptides was analyzed without further treatment. As each protein yielded a single peptide, the resultant mixture had the same level of complexity as the initial proteome sample (**Box 1**).

The previously published protocol, although effective, required multiple peptide-purification steps to separate the peptides from the excess reagents that were used, which had the consequence of reducing the yield of material. Hence there was a need for an enhanced methodology that minimized the processing steps in order to maximize the yield. We have therefore developed an improved strategy, which is described here. The most significant enhancement to the method was the elimination of the

BOX 1 | REMOVAL OF INTERNAL PEPTIDES BY BIOTINYLATION AND STREPTAVIDIN PURIFICATION

1. Using a C18 ZipTip (following the manufacturer's instructions), desalt a small portion (10 µl) of the digested peptide mixture and elute into 10 µl elution solution.
2. Add 40 µl of 20 mM phosphate buffer (20 mM Na_2HPO_4 (pH 7.5)) to give a total volume of 50 µl.

▲ CRITICAL STEP Do not use an amine-containing buffer (e.g., Tris or ammonium bicarbonate). The free amines will quench the NHS-biotin reagent and prevent biotinylation of peptides.

3. Reconstitute 1 mg EZ-Link NHS biotin in 50 µl DMF.

▲ CRITICAL STEP EZ-Link NHS biotin should be prepared immediately before use, and the solution should not be stored and reused.

4. Add 1 µl biotin solution to the desalted peptide mixture and incubate overnight at 4 °C.
5. Using a C18 ZipTip, desalt the entire biotinylated peptide mixture and elute into 10 µl elution solution.
6. Dilute the biotinylated peptides in 10 µl binding buffer (20 mM Na_2HPO_4 and 0.15 M NaCl (pH 7.5)) to give a total volume of 20 µl.
7. Remove 20 µl streptavidin Sepharose from the stock preparation in ethanol and pipette into a 0.5 ml microcentrifuge tube.
8. Remove excess ethanol and wash the streptavidin Sepharose in 100 µl (five volumes) of binding buffer.
9. Centrifuge the Sepharose at 2,000*g* at room temperature for 20 s, remove the binding buffer and discard.
10. Add the desalted biotinylated peptide mixture, vortex and incubate with turning for 4 h at room temperature.
11. Centrifuge the Sepharose/peptide mixture at 2,000*g* at room temperature for 20 s, remove and retain supernatant (peptide mixture).
12. Proceed to mass-spectrometric analysis.

biotinylation step. An initial acetylation step is still required to block N-terminal peptides, but instead of targeting internal peptides by biotinylation and removing them with streptavidin we use a commercially available amine reactive immobilized reagent (NHS-activated Sepharose) to react and retain internal peptides in one step. NHS-Sepharose is efficient both in respect of amine binding and subsequent leakage of bound amines[17]. The peptide mixture is incubated with the NHS-activated Sepharose until coupling is complete, the Sepharose beads are then removed by brief centrifugation, and the unbound fraction is removed and analyzed without further treatment (**Fig. 2**).

In brief, samples are exchanged into a compatible (non-amine containing) buffer prior to acetylation — if the protein mixture can be generated directly in such a buffer this step is not required. After excess acetylation reagent is removed, the proteins are concentrated by acid precipitation prior to digestion with trypsin or another protease. The proteolyzed mixture is then exposed to activated NHS Sepharose, which removes all peptides with proteolytically exposed amino groups — in other words, all internal peptides (**Fig. 3**). To illustrate the method, we use soluble proteins from *Escherichia coli*, but this approach is amenable to any protein or protein mixture. Indeed, the soluble protein preparation from chicken skeletal muscle (obtained frozen from a local supermarket) has the advantage of a considerable bias in protein expression[18,19], and should generate an N-terminal preparation dominated by a few peptides from abundant proteins, which is readily assessed by matrix-assisted laser desorption ionization-time of flight (MALDI-ToF) MS.

We prefer to use acetylation as the amino-blocking reagent, although there are many other reagents that could be used. Many proteins are naturally acetylated at their N terminus, and by using the same chemical modification, we are effectively coalescing the N-terminal preparation into the same analytical space. It is straightforward to discriminate between naturally and chemically acetylated N-termini by the use of [$^3H_6$] acetic anhydride, which reacts with primary amines and introduces a mass shift of +45 Da instead of +42 Da. If modifications other than acetylation are used, it introduces the possibility of further resolving the N-terminal peptide mixture by, for example, affinity capture or a shift in hydrophobicity[11].

There is concern within the field of proteomics over the value of 'one hit wonders'[20], whereby a protein is identified using data derived from a single peptide. Part of this concern relates to the lack of information regarding the location of the peptide within the parent protein, and also to the search space required for identification (the entire database of candidate peptides must be given equal validity in the analysis). The strategy we have developed overcomes the stigma associated with 'one hit wonders', by anchoring the peptides at a precise location within the parent protein; it is therefore possible to limit the database search to a small subset of peptides.

The bias in the preparation of N-terminal peptides has consequences for the subsequent database search. Two strategies are used.

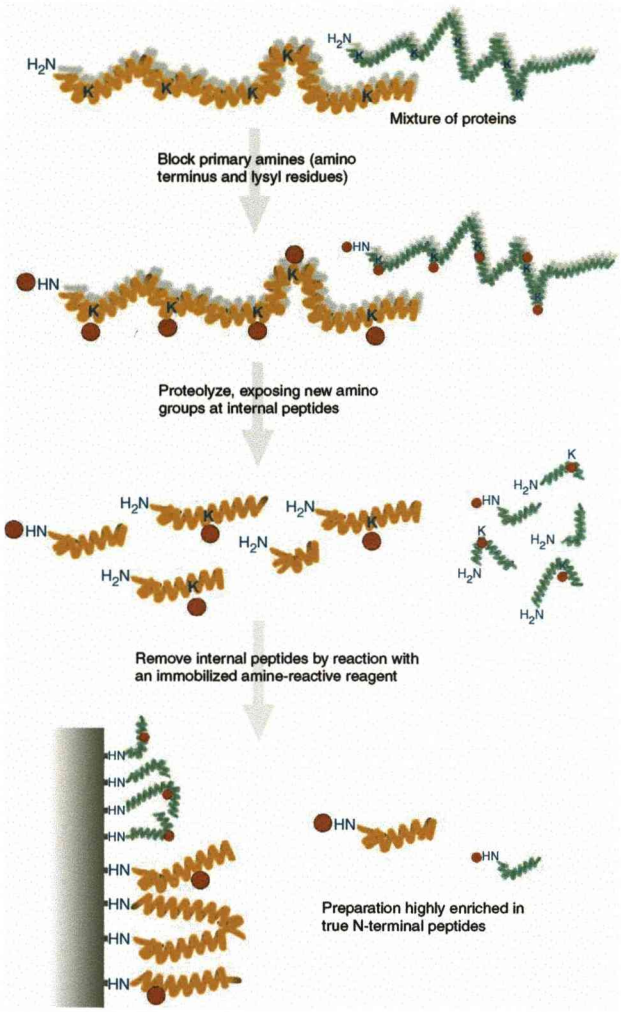


Figure 2 | Scheme showing the chemistry involved in N-terminal purification. Free amino groups (α and ε) are acetylated prior to proteolysis, which results in a mixture of N-terminally acetylated (true N-terminal) and non-acetylated (internal) peptides. Subsequent incubation of the peptide mixture with an immobilized amine-reactive reagent creates a preparation enriched in N-terminal peptides.

First, an entire database is searched, in which case the analyte peptides will be identified as high quality 'hits' that are also true N termini. However, this is predicated on the availability of knowledge of the true N terminus of each database entry. The complexity of post-translational processing and the fact that most proteomics databases are derived from cDNA or genomic data, creates a need for new search algorithms, yet to be developed, which capitalize on the drastically reduced search space and the positional bias. It is relevant to note that the preparation of N-terminal peptides can be employed to define the true N-termini of many proteins, and is therefore a useful tool in defining the true proteome.

MATERIALS

REAGENTS

- HPLC grade water
- HPLC grade acetonitrile (ACN) **! CAUTION** Flammable
- Luria broth (LB; Merck)
- Bugbuster protein extraction reagent (BB; Novagen-EMD Biosciences, cat. no. 70584) or any equivalent bacterial protein extraction reagent

- Coomassie Plus® protein assay (Perbio Science)
- Acetylation reagent: sulfo-NHS acetate (Pierce, cat. no. 26777)
- Acetylation buffer: 20 mM sodium carbonate (Na_2CO_3), pH 8.5, or other non-amine-containing buffers, such as phosphate or HEPES (pH 7–9)
  ▲ CRITICAL Do not use an amine-containing buffer (e.g., Tris or ammonium bicarbonate) as the free amines will quench the reagent and prevent protein acetylation
- Quenching reagent: Tris(2-aminoethyl)amine, polymer bound (Sigma, cat. no. 472107)
- 1 mM HCl
- Trichloroacetic acid (TCA) ! CAUTION Causes severe burns
- Trifluoroacetic acid (TFA) ! CAUTION Causes severe burns
- Diethyl ether ! CAUTION Highly flammable
- Digestion buffer: 20 mM Na_2HPO_4, pH 7.5
- Trypsin, sequencing grade (Roche, cat. no. 11 478 475 001) or any other proteolytic enzyme
- EZ-Link NHS-biotin (Pierce, cat. no. 20217)
- Dimethylformamide (DMF) ! CAUTION Harmful by inhalation, ingestion or skin contact
- Streptavidin Sepharose™, High Performance (GE Healthcare, cat. no. 17-5113-01)
- NHS-activated Sepharose™ 4 Fast Flow, stored in propanol (GE Healthcare, cat. no. 17-0906-01)
- Binding buffer: 20 mM Na_2HPO_4 and 0.15 M NaCl (pH 7.5)
- MALDI matrix: α-cyano-4-hydroxycinnamic acid (CHCA; Sigma, cat. no. C2020) ! CAUTION Irritating to eyes, respiratory system and skin
- Reverse-phase running buffer (A) 0.1% formic acid
- Reverse-phase eluting buffer (B) 90% ACN:0.1% formic acid
- Bacterial strains: *E. coli* BL21 ëDE3 (any other commercially available laboratory strain is suitable); frozen competent cells can be obtained from many suppliers, including Stratagene, Promega and Genlanatis
  ! CAUTION Use good microbiological practice in handling and disposing of this *E. coli* laboratory strain
- Chicken muscle (we used skeletal muscle tissue from *Gallus gallus* obtained frozen from a local supermarket)

EQUIPMENT

- Slide-A-Lyzer® dialysis cassettes, 500 µl to 3 ml, 10,000 molecular weight cut-off (Pierce, cat. no. 66425)
- ZipTip C18 pipette tips (Millipore, cat. no. ZTC18S008)
- Standard spectrophotometer for absorbance readings in the visible range, including 600 nm
- 1.5 and 0.5 ml plain microcentrifuge tubes
- Homogenizer
- Reverse-phase column: C18 3 µm particle size (100), 75 µm diameter × 150 mm long (Dionex)
- MALDI-ToF mass spectrometer (Waters MALDI-R, Shimadzu Axima TOF2 or equivalent)
- Electrospray-ionization tandem mass spectrometer coupled to a high-resolution nanoflow chromatography system (Dionex 3000 coupled to a Thermo Finnigan LTQ or other tandem mass spectrometer)

REAGENT SETUP

LB Dissolve 25 g LB powder in 1 l distilled water. The pH should be 7.0 ±0.2 at 25 °C; if not, adjust with HCl or NaOH as appropriate. Sterilize by autoclaving for 15 min at 121 °C.

Matrix for MALDI-ToF MS Prepare 50 ml of 50% (vol/vol) acetonitrile-0.1% (vol/vol) TFA; store at room temperature (20–25 °C). Prepare a fresh saturated solution of ~10mg CHCA in 1ml 50% (vol/vol) ACN-0.1% (vol/vol) TFA.

NHS-activated Sepharose Centrifuge the NHS-Sepharose slurry at 2,000*g* at room temperature for 20 s and remove excess propanol from the beads Wash the beads in five volumes of cold 1 mM HCl, vortex and remove HCl by centrifugation (as before). Wash in two volumes of binding buffer, remove by centrifugation. ▲ CRITICAL In order to retain maximum binding capacity of the pre-activated medium prior to the coupling step, use cold (0–4 °C) solutions. The time interval for all washing steps must be minimized. Prepare all required solutions prior to coupling ligand.

Chicken muscle Homogenize 0.5 g chicken skeletal muscle in 5 ml acetylation buffer. Centrifuge for 45 min at 13,000*g* at 4 °C, remove the supernatant fraction and use immediately or store at –20 °C. Determine the protein concentration of the soluble fraction using the Coomassie Plus® protein assay.

***E. coli* cell lysate** Using a single colony of *E. coli*, inoculate 10 ml LB and incubate overnight at 37 °C with shaking. Transfer 500 µl of the overnight culture to 50 ml prewarmed (to 37 °C) fresh LB media (1:100 dilution) and incubate the culture with shaking. Remove 1 ml samples at hourly intervals and determine the absorbance at 600 nm. Monitor growth rate until the early stationary phase is reached. Transfer the culture to a pre-weighed 50 ml centrifuge tube and centrifuge at 1,200*g* for 10 min at 4 °C. Decant the supernatant and weigh the tube again to determine the wet weight of the cell pellet. For ≤1 g of wet cell pellet, add 2.5 ml BB and, to ensure good resuspension, place cells on a rocker platform at room temperature for 15 min. Centrifuge the cells at 16,000*g* at room temperature for 20 min. Set the braking speed at low to give a gentle rotor deceleration. Remove the supernatant (soluble fraction) and use immediately or store at –20 °C. Determine the protein concentration of the soluble fraction using the Coomassie Plus® Protein assay.

EQUIPMENT SETUP

Slide-A-Lyzer® cassettes Before using the cassettes check for leaks by injecting the maximum amount of distilled water (3 ml) into each. Inject water through one of the four valves located in each corner of the cassette. Each valve can be used only once, so mark the cassette using a pen when a valve has been used. Attach a float to the top of the cassette and place into 1 litre acetylation buffer for 20 s in order to wet the membrane. Dry off the cassette by tapping it lightly onto a paper towel.

HPLC reverse-phase gradient The reversed phase chromatography was conducted at a flow rate of 0.3 µl min^{-1} over 3 h. The three phases of the gradients were as follows: 0–140 min, 0–50% buffer B (linear); 140–160 min, 50% buffer to 80% buffer B (linear); and 160–180 min, 80% buffer B to 0% buffer B.

PROCEDURE

Dialysis of *E. coli* proteins into acetylation buffer ● TIMING 4 h

1| Dialyze *E. coli* cell lysate into acetylation buffer using Slide-A-Lyzer® cassettes (see EQUIPMENT SETUP). Inject 3 ml *E. coli* cell lysate into the cassette through an unused valve. Attach the float and place into 1 l acetylation buffer. Leave to dialyze for 4 h at room temperature with stirring.

■ **PAUSE POINT** Dialyzed samples can be used immediately or stored for several months at –20 °C.

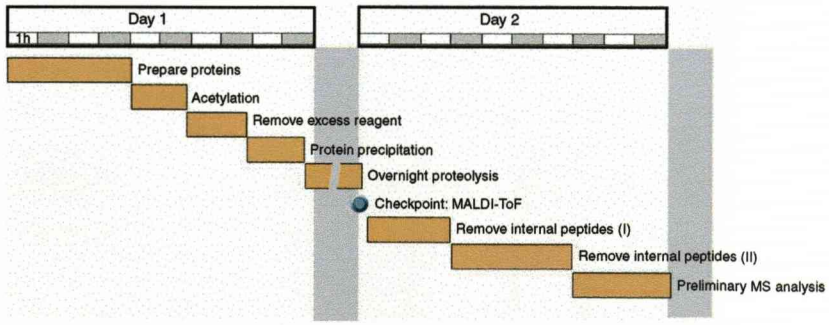


Figure 3 | Gantt chart for a typical N-terminal peptide-purification experiment. The entire procedure takes <2 d.

2| Determine the protein concentration of the dialyzed soluble fraction using the Coomassie Plus® protein assay. The protein concentration should be in the range of 1 to 5 μg μl^{-1}.

Acetylation of intact proteins ● TIMING 2 h

3| Reconstitute 1 mg sulfo-NHS acetate into 50 μl acetylation buffer.

▲ **CRITICAL STEP** Reconstitute sulfo-NHS acetate immediately before use. The NHS-ester readily hydrolyses and becomes unreactive.

4| Add 50 μl (1 mg) of the reconstituted acetylation reagent to 50 μg of the protein recovered from the dialyzed *E. coli* cell lysate (or protein mixture of choice). Incubate at room temperature for 2 h.

Removal of excess acetylation reagent ● TIMING 1 h

5| Add ~5 mg quenching reagent (Tris(2-aminoethyl)amine, polymer bound), vortex for 1 min and incubate with gentle agitation for 1 h.

▲ **CRITICAL STEP** This treatment has a major influence on the overall success of the process, and obviates the addition of free amines to inactivate excess reagent, which would then have to be removed before proceeding.

6| Remove amine-scavenging beads by filtration or centrifugation.

Precipitation of acetylated proteins ● TIMING ~1.5 h

7| Add 600 μl (five volumes) of cold 30% TCA to the protein mixture, vortex and incubate on ice for 1 h.

8| Centrifuge at 13,000*g* at room temperature for 2 min to pellet the protein.

9| Carefully remove the TCA supernatant fraction from the pellet and discard.

10| Add 200 μl diethyl ether to the pellet and agitate using a pipette tip.

! **CAUTION** Use a fume hood when pipetting ether.

11| Centrifuge for 10 s at 13,000*g* at room temperature.

12| Repeat Steps 10 and 11 twice more (three ether washes in total).

13| Remove diethyl ether and place tube at 37 °C for 5 min with the lid open to evaporate the excess.

Proteolysis ● TIMING overnight

14| Resuspend the diethyl ether-washed protein pellet in 50 μl digestion buffer. Digest overnight at 37 °C with 1 μg trypsin (or any other proteolytic enzyme; 1:50 enzyme:substrate).

■ **PAUSE POINT** Once digested, acetylated peptides can be stored for a few months at −20 °C.

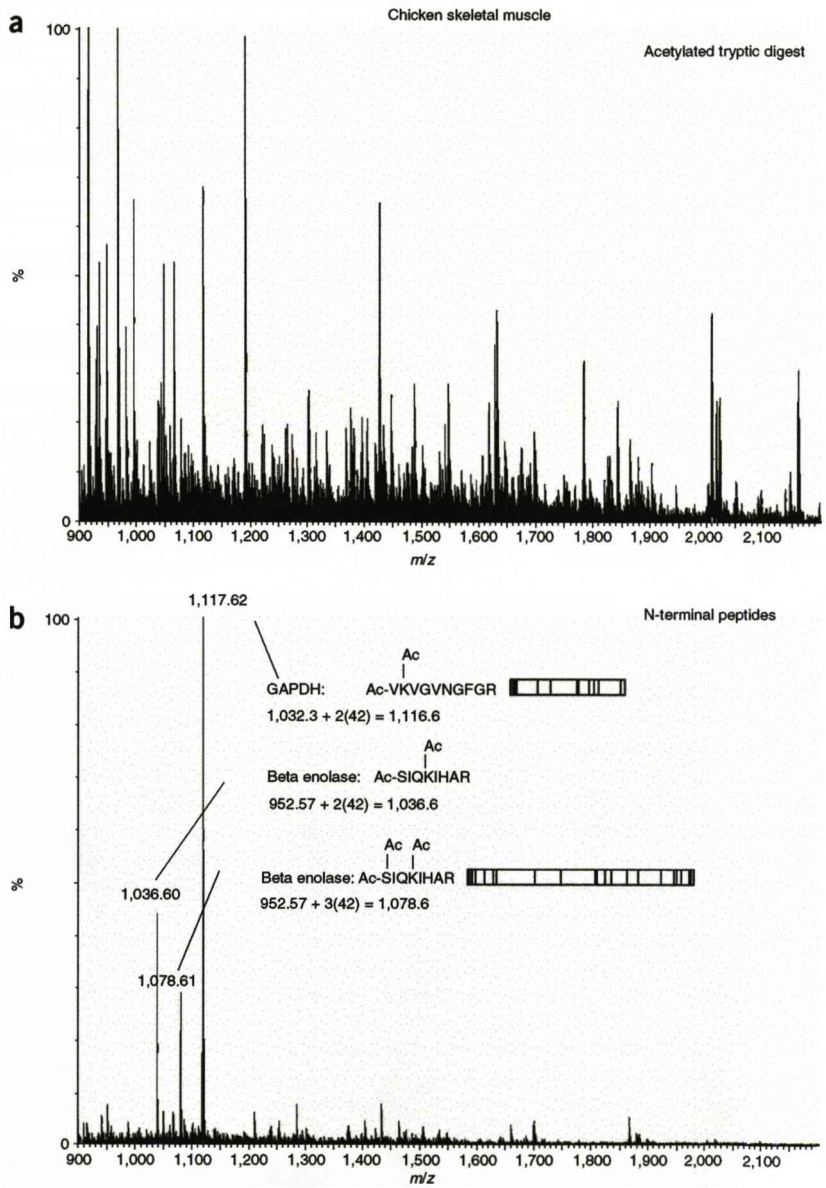


Figure 4 | Isolation of N-terminal peptides from chicken skeletal muscle soluble fraction. The starting material is a complex mixture of proteins, which simplifies to a relatively straightforward N-terminal peptide mix. (**a**) The entire tryptic digest of acetylated proteins. (**b**) Following application of the simplification protocol, the major ions labeled correspond to the N-terminal peptides from some of the most abundant soluble proteins in skeletal muscle.

MALDI-ToF analysis ● TIMING ~1 h

15| At this stage in the protocol, it is good practice to monitor acetylation and proteolysis of the digested peptides by MALDI-ToF using a Waters MALDI-R or any equivalent MALDI-ToF mass spectrometer. Prepare samples as follows: (i) dilute sample 1:20 in matrix (2 µl sample + 38 µl matrix); (ii) pipette 1 µl onto a clean MALDI target and allow to air dry; (iii) acquire data over the range of 900 to 3,500 *m/z*. At this stage of the protocol, a complex spectrum should be observed (corresponding to peptides derived from the entire protein mixture). The level of acetylation is difficult to determine; however, the observation of ArgC (as opposed to tryptic) peptides is a good indication that a sufficient degree of modification has occurred. Internal peptides can now be removed using NHS-activated Sepharose (continue to Step 16) or by biotinylation and streptavidin purification (**Box 1**).

Coupling of internal peptides to NHS-activated Sepharose ● TIMING ~24 h

16| Dilute the digested peptides in 50 µl binding buffer (20 mM Na_2HPO_4 and 0.15 M NaCl (pH 7.5)).

17| Remove 100 µl NHS-activated Sepharose from the stock preparation in propanol, and pipette into a 1.5-ml microcentrifuge tube. Wash as described in the REAGENT SETUP.

18| Add acetylated peptides (50 µg) to NHS-Sepharose, vortex and incubate with turning for 4 h at room temperature.

19| Centrifuge the Sepharose/peptide mixture at 2,000*g* at room temperature for 20 s, then remove and retain the supernatant (peptide mixture).

20| Prepare a second aliquot (100 µl) of NHS-activated Sepharose (repeat Step 17) and add the peptide mixture.

21| Incubate overnight at 4 °C with turning.

▲ **CRITICAL STEP** A second incubation with NHS Sepharose is necessary for complete coupling of peptides to Sepharose and to minimize leakiness of the procedure, wherein internal peptides can appear in the N-terminal peptide preparation.

22| Remove the peptide mixture from NHS-Sepharose (as before), divide into 10 µl aliquots and proceed to mass-spectrometric analysis.

■ **PAUSE POINT** The peptide mixture can be stored at −20 °C prior to MS.

MALDI-ToF analysis ● TIMING ~1 h

23| Prepare samples for MALDI-ToF analysis as described in Step 15.

Tandem mass-spectrometric analysis ● TIMING ~3 h for one sample

24| Separate the samples using a microcapillary reverse-phase column

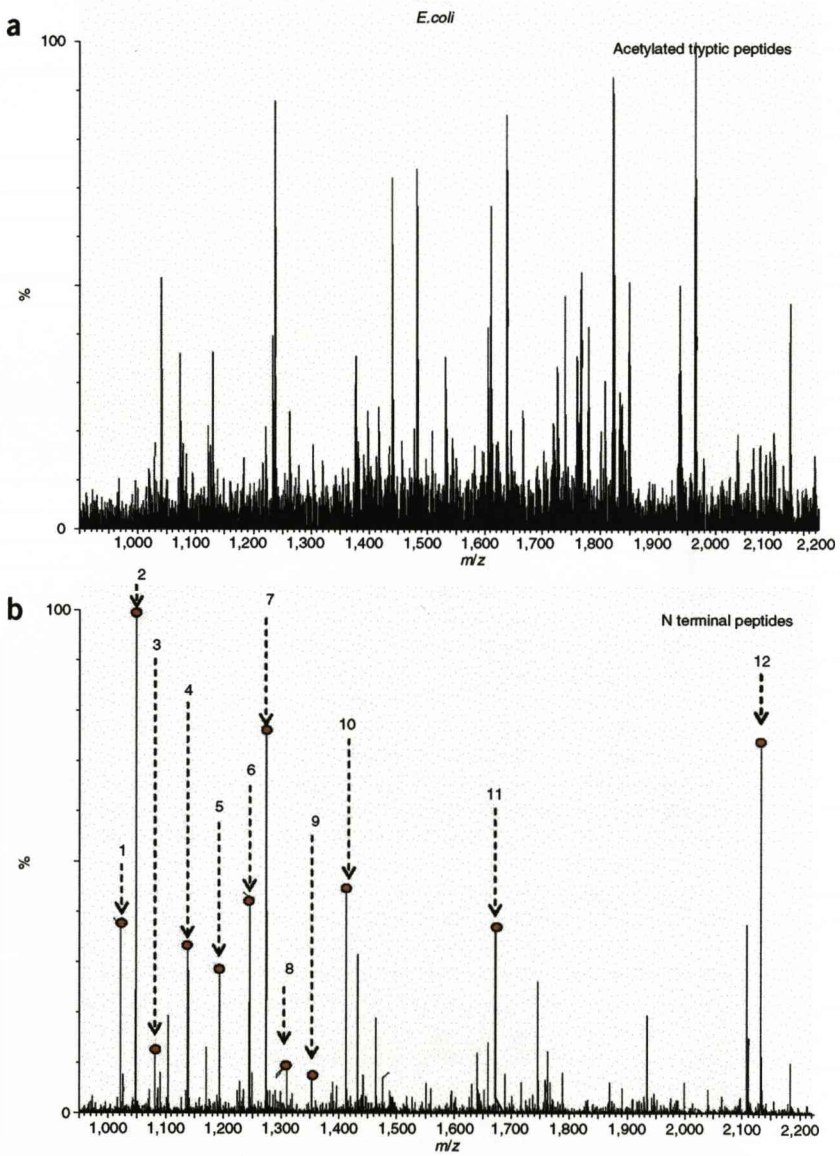


Figure 5 | Isolation of N-terminal peptides from *E. coli* cell lysate. (**a**) The entire tryptic digest of acetylated proteins. (**b**) Following application of the simplification protocol, the major ions labeled correspond to N-terminal peptides from some of the most abundant proteins in the sample.

(using reverse phase running buffer (A) and reverse phase eluting buffer (B)) in line with an eletrospray-ionization ion-trap tandem mass spectrometer.

Data analysis

25| Search the tandem mass spectra against Swiss-Prot using MASCOT or the TurboSequest program (ThermoElectron). Database search parameters include the fixed modifications of N-terminal acetylation and lysine acetylation, and the variable modification of O-acetylated serine (this is rare, however, so searches should be conducted both with and without this modification).

● **TIMING**

Dialysis of *E. coli* proteins into acetylation buffer: 4 h
Acetylation of intact proteins: 2 h
Removal of excess acetylation reagent: 1 h
Precipitation of acetylated proteins: ~1.5 h
Proteolysis: overnight
MALDI-ToF analysis: ~1 h
Coupling of internal peptides to NHS-activated Sepharose: ~24 h
MALDI-ToF analysis: ~1 h
Tandem mass-spectrometric analysis: ~3 h for one sample

? TROUBLESHOOTING

See **Table 1.**

TABLE 1 | Troubleshooting for the preparation of N-terminal peptides.

Problem	Possible cause	Solutions
Incomplete acetylation	Instability of acetylation regent (sulfo-NHS acetate and acetic anhydride)	Prepare sulfo-NHS acetate freshly and use immediately
	Acidification of mixture by excess acetic anhydride lowers pH	Use an acetylation buffer that is able to maintain the pH at ~8.5 for the duration of the reaction
	Accessibility of some primary amino groups is impaired by the 3D structure of the protein	Use buffers that denature or partially unfold the protein, or increase the time for acetylation (e.g., chaotropic buffers and/or reducing agents)
	Amines present in sample buffer compete with proteins for acetylation	Ensure that the buffer used does not contain amines
Incomplete proteolysis	Precipitated proteins not washed adequately to remove all residual trichloroacetic acid, which will acidify the digestion buffer and shift the pH down from the optimal value for trypsin action	Wash protein pellets carefully with ether to remove all TCA, or precipitate proteins with acetone; TCA will denature the protein mixture more effectively and increase the efficiency of proteolysis
Incomplete removal of peptides, identified by database searching.	Insufficient NHS-activated Sepharose used in the coupling step	Use a larger excess of NHS-activated Sepharose; ensure the NHS-Sepharose is freshly prepared
	Not enough time allowed for coupling	Use longer coupling times; repeat the reaction with a second batch of freshly washed NHS-Sepharose
	pH of coupling reactions not optimal	Ensure the system is buffered effectively at pH 8.0 ± 0.5; ensure that no amine-containing buffers, including Tris, have been added to the mixture; make sure that the internal peptide is not a true intracellular N terminus generated by endogenous proteolysis or by ectopic digestion after tissue breakage
Poor MALDI spectra	Phosphate buffer/PBS affecting ionization	Dilute the sample further or desalt using a C18 column; change the proteolysis buffer to a different system, such as HEPES
	Sample too complex for meaningful interpretation	This will be normal for a complex protein mixture in which no proteins predominate; separate the sample using HPLC (as in Step 24) and analyze fractions by MALDI, or move directly to LC-MS/MS or LC2-MS/MS (MUDPIT)

TABLE 2 | *E. coli* N-terminal peptides from MALDI spectrum.

Spot	Protein	Mass (Da)	Sequence
1	UPF0304 protein yfbU (P0A8W8)	1,021.43	MEMTNAQR
2	Elongation factor Tu (EF-Tu) (P0A6N1)	1,048.52	SKEKFER
3	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase (P62707)	1,081.69	AVTKLVLVR
4	Enolase (P0A6P9)	1,138.71	SKIVKIIGR
5	10 kDa chaperonin (P0A6G1)	1,192.61	MNIRPLHDR
6	Glyceraldehyde-3-phosphate dehydrogenase A (P0A9B4)	1,244.69	TIKVGINGFGR
7	Copper-resistance protein D (Q47455)	1,274.67	MNDLIMIVIR
8	D,D-heptose 1,7-bisphosphate phosphatase (Q8FKZ1)	1,299.72	AKSVPAIFLDR
9	Putative HTH-type transcriptional regulator yeaT (P76250)	1,353.71	MNNLPLLNDLR
10	Elongation factor Ts (EF-Ts) (P0A6P1)	1,412.79	AEITASLVKELR
11	Phosphoglycerate kinase (P0A799)	1,671.89	SVIKMTDLDLAGKR
12	β-lactamase (P62593)	2,132.30	HPETLVKVKDAEDQLQR

These ions should be visible even in a MALDI-ToF spectrum of an entire N-terminal preparation of *E.coli* soluble proteins.

ANTICIPATED RESULTS

MALDI spectra of a digest of acetylated proteins should yield a complex mass spectrum, which represents the most abundant peptides in the sample. This analysis serves as a 'check point' in which the extent of acetylation can be monitored, provided that a few key peptides can be recognized. Due to the complexity of the peptide mixture, it is difficult to assign peptides to individual peaks at this stage in the protocol. Following abstraction of internal peptides to NHS-activated Sepharose, the N-terminally enriched supernatant should produce a notably simpler mass spectrum. At this stage, it might be possible to assign the highest intensity signals to true N-terminal peptides.

Skeletal muscle soluble protein preparations are dominated by ~10–20 major proteins that are predominantly glycolytic enzymes[19,20]. This preparation has the advantage of generating relatively simple spectra, and is a valuable test system with which to practice the method. **Figures 4a** and **5a** show the entire tryptic digests for the acetylated chicken muscle and *E. coli* cell lysate samples. As expected the spectra are complex and it is not possible to identify any N-terminal peptides. However, **Figures 4b** and **5b** represent the unbound fraction for the two samples, which should be substantially N-terminally enriched. These elicit much simpler mass spectra, and it is possible to assign identities to the most intense signals (*E. coli* MALDI-ToF peak assignments are listed in **Table 2**). Due to the dynamic range of the skeletal muscle proteome, the N-terminal spectrum for chicken muscle is substantially less complex than the *E. coli* sample. The high abundance of glycolytic proteins found in skeletal muscle means that the lower-abundance proteins are not visible at this stage of analysis.

Using a 3-h HPLC gradient, tandem mass-spectrometric analysis should provide data on hundreds of N-terminal peptides (depending on sample complexity). When analyzed in this way, the *E. coli* N-terminally enriched preparation yielded >300 protein identifications. All identifications were from a search of the entire SwissProt database of *E. coli* proteins, and relatively few peptides were identified as internal sequences.

COMPETING INTERESTS STATEMENT The authors declare that they have no competing financial interests.

ACKNOWLEDGMENTS This work was supported by the Engineering and Physical Sciences Research Council.

Published online at http://www.natureprotocols.com
Rights and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Bogdanov, B. & Smith, R.D. Proteomics by FTICR mass spectrometry: top down and bottom up. *Mass Spectrom. Rev.* **24**, 168–200 (2005).
2. Wu, C.C. & MacCoss, M.J. Shotgun proteomics: tools for the analysis of complex biological systems. *Curr. Opin. Mol. Ther.* **4**, 242–250 (2002).
3. Liu, H., Lin, D. & Yates, J.R. Multidimensional separations for protein/peptide analysis in the post-genomic era. *Biotechniques* **4**, 898–902 (2002).
4. Mirzaei, H. & Regnier, F. Structure specific chromatographic selection in targeted proteomics. *J. Chromatogr. B* **817**, 23–34 (2005).
5. Smolka, M.B., Zhou, H., Purkayastha, S. & Aebersold, R. Optimization of the isotope-coded affinity tag-labeling procedure for quantitative proteome analysis. *Anal Biochem.* **297**, 25–31 (2001).
6. Raggiaschi, R., Gotta, S. & Terstappen, G. Phosphoproteome analysis. *Biosci. Rep.* **25**, 33–44 (2005).
7. Kweon, H.K. & Hakansson, K. Selective zirconium dioxide-based enrichment of phosphorylated peptides for mass spectrometric analysis. *Anal. Chem.* **78**, 1743–1749 (2006).
8. West, I. & Goldring, O. Lectin affinity chromatography. *Methods Mol. Biol.* **59**, 177–185 (1996).
9. Gevaert, K., Van Damme, P., Martens, L. & Vandekerckhove, J. Diagonal reverse-phase chromatography applications in peptide-centric proteomics: ahead of catalogue-omics? *Anal. Biochem.* **345**, 18–29 (2005).
10. Martens, L. *et al.* The human platelet proteome mapped by peptide-centric proteomics: a functional protein profile. *Proteomics* **5**, 3193–3204 (2005).
11. Gevaert, K. *et al.* Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat. Biotechnol.* **21**, 566–569 (2003).
12. Kasai, K.-I. Trypsin and affinity chromatography. *J Chromatogr. A* **597**, 3–18 (1992).
13. Sechi, S. & Chait, B.T. A method to define the carboxyl terminal of proteins. *Anal. Chem.* **72**, 3374–3378 (2000).
14. Yamaguchi, M. *et al.* High-throughput method for N-terminal sequencing of proteins by MALDI mass spectrometry. *Anal. Chem.* **77**, 645–651 (2005).
15. Chelius, D. & Shaler, T.A. Capture of peptides with N-terminal serine and threonine: a sequence-specific chemical method for peptide mixture simplification. *Bioconjugate Chem.* **14**, 205–211 (2003).

16. McDonald, L., Robertson, D.H., Hurst, J.L. & Beynon, R.J. Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides. *Nat. Methods* **2**, 955–957 (2005).
17. Van Sommeren, A.P.G., Machielsen, P.A.G.M. & Gribnau,, T.C.J. Comparison of three activated agaroses for use in affinity chromatography: effects on coupling performance and ligand leakage. *J. Chromatogr. A* **639**, 23–31 (1993).
18. Hayter, J.R., Robertson, D.H.L., Gaskell, S.J. & Beynon, R.J. Proteome analysis of intact proteins in complex mixtures. *Mol. Cell Proteomics* **2**, 85–95 (2003).
19. Doherty, M.K. *et al.* The proteome of chicken skeletal muscle: changes in soluble protein expression during growth in a layer strain. *Proteomics* **4**, 2082–2093 (2004).
20. Veenstra, T.D., Conrads, T.P. & Issaq, H.J. What to do with "one-hit wonders"? *Electrophoresis* **25**, 1278–1279 (2004).

research articles Journal of proteome research

Asparagine Deamidation and the Role of Higher Order Protein Structure

Jenny Rivers, Lucy McDonald, Ian J. Edwards, and Robert J. Beynon*

Proteomics and Functional Genomics Group, Faculty of Veterinary Science, University of Liverpool, Crown Street, Liverpool L69 7ZJ, United Kingdom

Received July 11, 2007

The 'protein world' exhibits additional complexity caused by post-translational modifications. One such process is nonenzymic deamidation of asparagine which is controlled partly by primary sequence, but also higher order protein structure. We have studied the deamidation of an N-terminal peptide in muscle glyceraldehyde 3-phosphate dehydrogenase to relate three-dimensional structure, proteolysis, and deamidation. This work has significant consequences for identification of proteins using peptide mass fingerprinting.

Keywords: Deamidation • proteolysis • protein structure • asparagine • aspartic acid • peptide mass fingerprinting

Introduction

The emergence of new analytical methods for protein characterization has led to the recognition that there is an additional dimension of complexity in the protein world created by a wide range of post-translational modifications. Some of these modifications are specific and are part of the obligatory maturation process of a protein, such as the removal of propeptides. Other changes are transient, reversible, and may only operate on a subset of molecules in the protein pool (the best understood is phosphorylation). Other irreversible changes, such as deamidation or lysine aldehyde mediated cross-linking, are nonenzymic, and the longevity of the protein may be reflected in the accumulation of such changes.

Deamidation of the side chain of asparagine residues is a nonenzymic process[1] (www.deamidation.org). The conversion of asparagine to aspartic acid or isoaspartic acid elicits a local change in charge, and has the potential to impose a self-timer on protein molecules, altering activity or stability with lifetime kinetics.[2-5] The ability to include a nonenzymic irreversible change into a protein that elicits a small steric change but a substantial local alteration in electrostatic potential could provide an opportunity to evolve a programmable irreversible change of state into a protein. Most studies on asparagine deamidation have been conducted with model peptides[6] which are essentially devoid of higher order structure and which permit the peptide backbone and side chain to adopt a conformation compatible with the cyclic intermediate that is required for this reaction to take place. Since the flexibility and conformational freedom of the peptide is modified by the nature of the amino acids, the rate of deamidation of model peptides is strongly influenced by the flanking residues[6] and the primary influence on the rate of asparagine deamidation is the amino acid C-terminal to the asparagine residue. From studies of model peptides, the highest rate of deamidation is obtained when the carboxyl neighbor is glycine, yielding a half-time for deamidation of around 24 h.[6] This is probably because the lack of a Cβ atom minimizes steric hindrance and permits ready formation of the five-membered imide conducive to the deamidation reaction. The N-terminal neighbor has a minor effect on the rate of deamidation.[6] While most of our understanding of rates of peptide deamidation has derived from short, model peptides, the same sequences, when incorporated into protein structures, might acquire a relatively immobile backbone trajectory that could constrain the sequence to either favor or disfavor deamidation.

The resolution of modern mass spectrometers used routinely in proteomic analyses permits ready resolution of the monoisotopic peptide–ion from the ^{13}C isotopomer variants, even at charge states of +2 or +3. At this level of resolution, a deamidation event (Asn → Asp) would be readily recognized, as it elicits a mass shift of +0.985 Da (-NH_2 = 16.03 to -OH = 17.01). In circumstances where a peptide exists as a mixture of the amide and cognate acid species, a complex mass spectrum would ensue that appears as an atypical isotopomer distribution for a peptide of that mass. It follows that partial deamidation events should be readily observed by examination of the atypical profile, particularly without prior chromatographic separation that would resolve the amide and cognate acid in chromatographic space.

In the course of proteomics studies of soluble proteins in skeletal muscle,[7] we observed that a peptide from one protein in particular exhibited a noticeable and atypical natural isotope distribution profile, consistent with a mixture of an asparagine-containing peptide and the cognate deamidation product. This peptide was derived from the N-terminus of an abundant protein, glyceraldehyde-3-phosphate dehydrogenase (GAPDH). We present here a comprehensive analysis that confirms that the 'atypical' isotope profile is in fact attributable to partial deamidation of an asparagine residue. Deamidation of -AsnGly-

* To whom correspondence should be addressed. Phone: +44 151 794 4312. Fax:+44 151 794 4243. E-mail: r.beynon@liv.ac.uk.

10.1021/pr070425l CCC: $40.75

Published on Web 02/05/2008

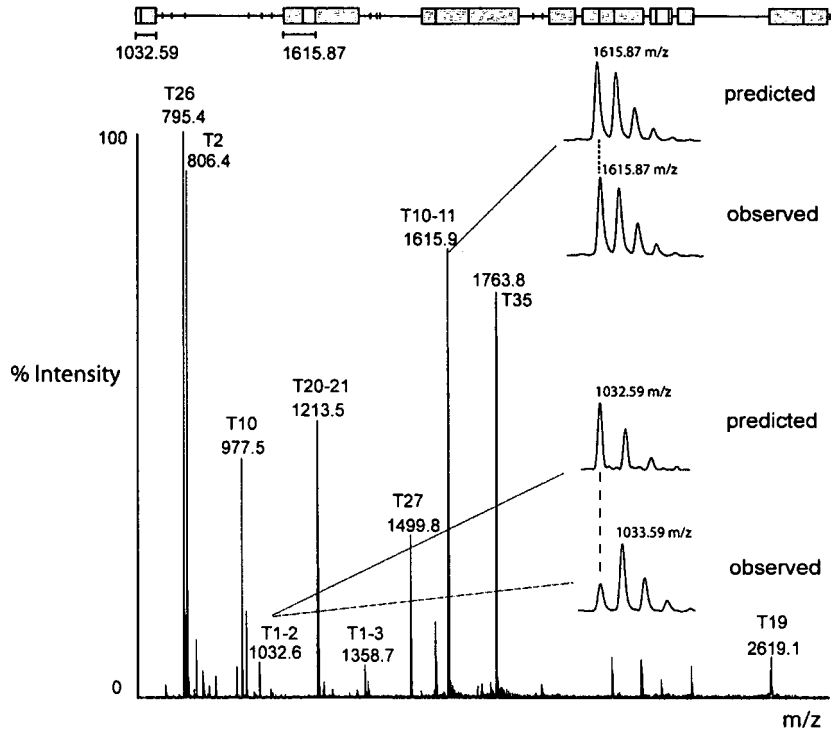


Figure 1. Atypical peptide mass spectrum consistent with deamidation. Glyceraldehyde 3-phosphate dehydrogenase (GAPDH; 1 mg/mL diluted to 0.2 mg/mL with 50 mM ammonium bicarbonate) purified from rabbit skeletal muscle (Sigma, Dorset, U.K.) was digested in solution with trypsin at a substrate/protease ratio of 100:1 by weight, and the masses of the resultant tryptic peptides were assessed by MALDI-ToF mass spectrometry; a coverage map is included at the top of the figure, with identified peptides indicated by a shaded block and those identified as part of a missed cleavage by an open block. The spectrum of a typical partial cleavage tryptic peptide (T10–11, *m/z* 1615.9) was compared with the mass spectrum predicted by the MS-Isotope tool (http://prospector.ucsf.edu/). This behavior, common to almost all other peptides, emphasized the atypical profile observed for the N-terminal partial cleavage peptide (T1–2, *m/z* 1032.6).

sequences occurs during sample preparation in proteomics,[8] and proteolysis conducted at lower pH and temperature will minimize artifactual deamidation.[9] Here, we show that deamidation is constrained by higher order structure and is enhanced after release of that conformational restraint by proteolysis. This observation has significance for the identification of deamidation events by protein or peptide mass spectrometry[10–12] and reinforces the role that protein conformation can play in this process.

Experimental Section

Materials and Reagents. Trypsin (sequence grade) was obtained from Roche Diagnostics (Lewes, U.K.). All other chemicals and solvents (HPLC grade) were purchased from Sigma-Aldrich Company Ltd. (Dorset, U.K.) and VWR International Laboratory Supplies (Leicestershire, U.K.).

One-Dimensional Gel Electrophoresis (1DGE). Purified GAPDH from rabbit skeletal muscle (Sigma, Dorset, U.K.) (10 μg) was electrophoresed through a 12.5% polyacrylamide gel and visualized with Biosafe Coomassie Brilliant Blue stain (Bio-Rad, Hemel Hempstead, U.K.). Gels were destained with a 10% acetic acid 10% methanol solution.

In-Gel Trypsin Digestion. Gel plugs containing GAPDH (identification confirmed by MALDI-ToF MS, results not shown) were excised from 1D gels using a glass pipet and transferred to an Eppendorf tube. To each tube, 25 μL of 50 mM ammonium bicarbonate, pH 8.2, and 50% (v/v) acetonitrile (ACN) was added and incubated at 37 °C for 20 min. This process was repeated until all of the stain had been removed. The plugs were then washed in 50 mM ammonium bicarbonate, which was subsequently discarded. The gel was dehydrated using 5 μL of ACN, and incubation at 37 °C was resumed for 30 min. Once dry, the gel was rehydrated in 50 mM ammonium bicarbonate (9 μL) containing trypsin (1μL of 100 ng/μL trypsin stock reconstituted in 50 mM acetic acid), and digestion was allowed to continue overnight at 37 °C; the digestion was halted by the addition of 2 μL of formic acid.

MALDI-ToF Mass Spectrometry. Peptides were analyzed by MALDI-ToF (M@LDI; Waters, Manchester, U.K.) mass spectrometry. For this, 1 μL of digested material was mixed with an equal volume of α-cyano-hydroxycinnamic acid in 50% (v/v) ACN and 0.1% (v/v) trifluoroacetic acid. This was allowed to dry, and peptides were acquired over the range 900–3000 *m/z*. For each combined spectrum, 20–30 spectra were acquired (laser energy typically 30%) with 10 shots per spectrum and a laser firing rate of 5 Hz. Data were processed using MassLynx software to subtract background noise using polynomial order 10 with 40% of the data points below this polynomial and a tolerance of 0.01. Spectral data were also smoothed by performing two mean smooth operations with a window of three channels. To confirm the assumption that both acid and amide forms of the peptide ionize with equal signal response in MALDI-ToF MS, the synthetic peptide for the amide form was allowed to fully deamidate (by incubation at 37 °C) and mixed in a known ratio with asparagine-containing peptide in a strong acidic solution to prevent further deamidation. The signal response from the two variants was identical.

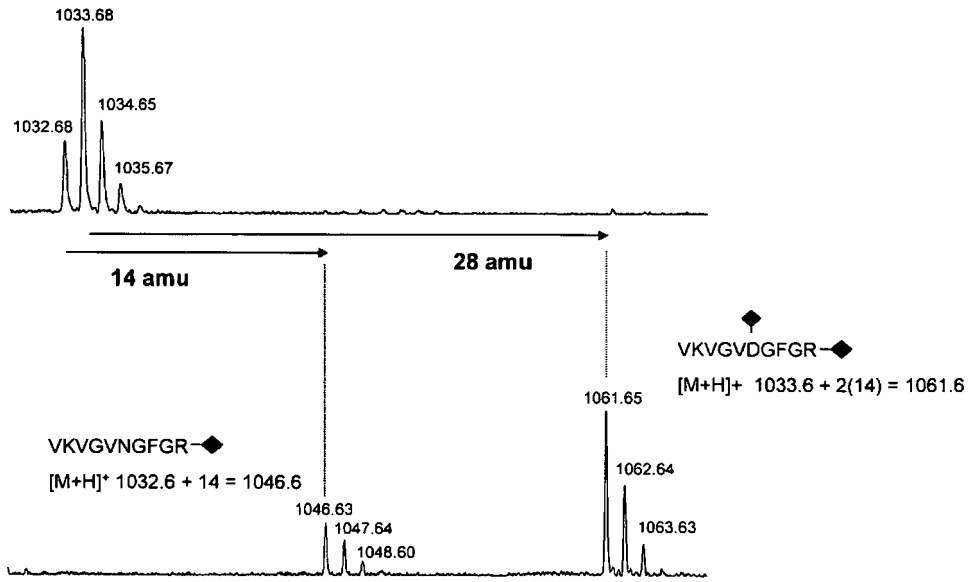


Figure 2. Esterification of acidic residues in the N-terminal peptide of GAPDH. Tryptic peptides recovered from an in-gel tryptic digest of GAPDH (purified from rabbit skeletal muscle, Sigma, Dorset, U.K.) were reacted with acetyl chloride and methanol to convert acidic residues to their corresponding methyl esters. The upper mass spectrum is the peptide resulting from partial deamidation of Asn_6, thus, is a mixture of two forms (asparagine containing and aspartic acid containing). The lower spectrum, obtained after esterification has resolved the peptide into two distinct reaction products at 1046.63 *m/z* and 1061.65 *m/z*, consistent with the addition of one and two methyl groups (+14.03 Da), respectively.

In-Solution Tryptic Digestion. Soluble protein (purified GAPDH; 1 mg/mL) was diluted 10-fold with 50 mM ammonium bicarbonate prior to addition of trypsin (100:1 substrate/protease). The reaction mixture was incubated at 37 °C for 24 h, and peptides were analyzed by MALDI-ToF MS.

Esterification of Peptides. A stock solution of methanol (1 mL, previously stored at −20 °C for 15 min) and acetyl chloride (150 μL) was prepared. An aliquot (10 μL) of this mixture was then added to a dried portion of the peptide pool recovered after in-gel digestion of the protein. The mixture was incubated at room temperature for 45 min prior to drying in a vacuum centrifuge. Esterified peptides were analyzed by MALDI-ToF MS.

Monitored Proteolysis of GAPDH. Digestion reaction mixtures with trypsin were stopped at selected time points after addition of enzyme by removing 10 μL and adding to an equal volume of 10% (v/v) formic acid. The fractions were subsequently stored at −20 °C until the end of the time course. Peptides were analyzed by MALDI-ToF MS.

Data Processing. The natural isotope profile for the acid VKVGVDGFGR and amide VKVGVNGFGR variants of the GAPDH N-terminal peptide were predicted using the MSIsotope tool provided online within the Protein Prospector Package (http://prospector.ucsf.edu/ucsfhtml4.0/msiso.htm). The intensities of each isotopomer peak were added, and the combined theoretical spectrum was compared with the intensities derived from the experimental mass spectrum. The sum of the squares of the deviation between predicted and experimental data was used to generate the object function, and the sole parameter (P_A) was the proportion of the acidic component (by definition, equal to $1 - P_N$, where P_N is the proportion of amide). The nonlinear optimization function (Solver) within Excel was used to obtain the best fit value of P_A. Additionally, some samples were analyzed by a high speed spectrum deconvolution tool, implemented as computer hardware in a field programmable gate array.[13]

Absolute Quantification of Proteolysis Using a Stable Isotope-Labeled Synthetic Peptide. The N-terminal peptide of GAPDH, of sequence VKVGVNGFGR and neutral mass 1041.59 Da, was synthesized by Sigma-Genosys (Dorset, U.K.) and was labeled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10 Da mass offset from the analyte peptide. For quantification of proteolysis, the synthetic peptide was added to digested material in 10% (v/v) formic acid to stop digestion and deamidation. Peptides were analyzed by MALDI-ToF MS, and the relative intensities of analyte peptide and internal standard were used to quantify the amount of peptide released from the protein during incubation with trypsin at 37 °C. As conversion of asparagine to aspartic acid alters the isotope envelope of the analyte peptide, the composite abundance of the entire isotopic envelope for both analyte and internal standard peptide was summed in each case. These data permitted the kinetics of proteolytic release of the N-terminal peptide from GAPDH to be calculated and were used along with the kinetics of deamidation to investigate the interaction between these two alternative processes. The rate of deamidation was measured across the time course of digestion by calculating the proportion of acid and amide variants of the peptide at each time point. This was done during proteolysis of GAPDH and for the synthetic peptide, at different temperatures.

Results and Discussion

One of the most abundant soluble sarcoplasmic proteins in skeletal muscle is glyceraldehyde 3-phosphate dehydrogenase, amounting to 11 ± 1% (mean ± SEM, $n = 3$) of soluble protein when resolved by 1D gel electrophoresis (1DGE) and analyzed by densitometry (data not shown) and up to 500 ± 50 nmol/g (mean ± SEM, $n = 4$) tissue when analyzed using the QconCAT method for absolute quantification.[13-15] MALDI-ToF spectra

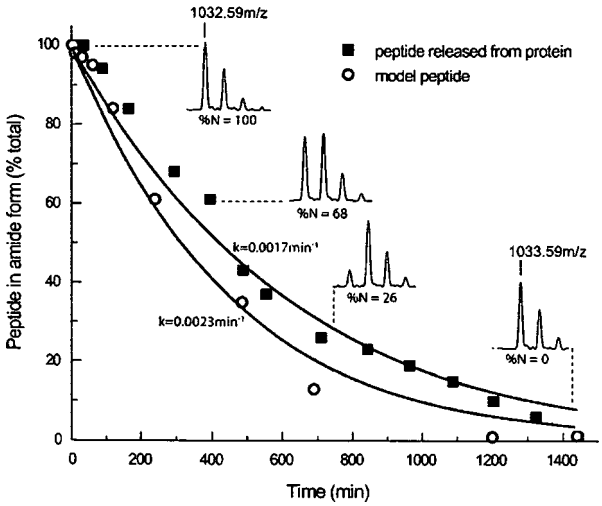


Figure 3. Time course of deamidation of the N-terminal peptide of GAPDH. Purified rabbit skeletal muscle GAPDH (Sigma, Dorset, U.K.; 1 mg/mL diluted to 0.2 mg/mL with 50 mM ammonium bicarbonate) was digested with trypsin (trypsin/protein 1:100) over 24 h at 37 °C. Proteolysis was stopped at 0, 2, 5, 10, 30, 60, 120, 240, 480, and 1440 min by mixing 10 μL from the digestion mixture with 10 μL of 10% (v/v) formic acid. The resulting peptides were analysed by MALDI-ToF mass spectrometry, and deamidation was monitored during proteolysis for the N-terminal peptide of sequence VKVGVNGFGR at 1032.59 *m/z*. The proportion of acid and amide variants was assessed as described in Experimental Section, from peak height data, and plotted as a function of time (closed squares). Peptide envelopes illustrating the conversion of acid to amide form in MALDI-ToF mass spectra corresponding to time points over 24 h are inserted above the data. To compare this with model peptide studies, the N-terminal peptide of GAPDH, of sequence VKVGVNGFGR and mass 1041.59 Da, was synthesised by Sigma-Genosys (Dorset, U.K.) and was labelled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10 Da mass offset from the analyte peptide. This peptide was incubated in 50 mM ammonium bicarbonate at 37 °C, and a sample of the peptide was added to an equal volume of 10% (v/v) formic acid at selected time points. The relative amounts of acid and amide variants of the peptide were measured using MALDI-ToF MS, and this was used to calculate the rate of deamidation. These data are presented as open circles. The solid lines are the trajectories taken by first-order decay for the synthetic peptide and the proteolyzed glyceraldehyde 3-phosphate dehydrogenase.

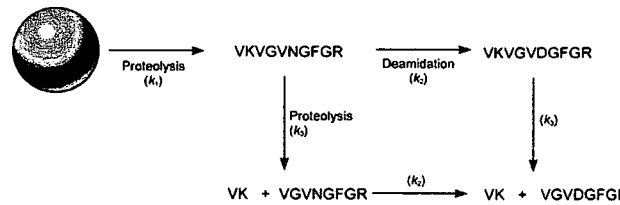


Figure 4. Model of proteolysis and deamidation of the N-terminal peptide of GAPDH. The simultaneous processes of proteolysis and release of the N-terminal peptide of GAPDH followed by deamidation of the asparagine residue to aspartic acid were modelled according to this scheme. The model also included the subsequent proteolysis of the N-terminal peptide (**VKVGVNGFGR** or **VKVGVDGFGR**) at the internal arginine residue to generate a dipeptide and a truncated peptide (**VK+VGVNGFGR** or **VK+VGVDGFGR**). In this scheme, we assumed that the rate of deamidation was the same, whether in the full length or truncated N-terminal peptide, and that the rate of removal of the N-terminal dipeptide was independent of the amide/acid variants.

for this protein, isolated by 1DGE and digested with trypsin prior to MS analysis are of high quality, give very high probability identification of this protein (not shown), and yield approximately 20 peptides, ranging from 805.5 *m/z* to 2265.4 *m/z*. Close inspection of each peptide indicated that for most, the observed mass isotopomer distribution was as expected, and was in close agreement to the distribution predicted by the MsIsotope program (http://prospector.ucsf.edu/). One peptide in particular (VKVGVNGFGR, $[M+H]^+$ 1032.58 *m/z*) was notably different from the others, inasmuch as the isotope distribution profile was far removed from the predicted profile (Figure 1). In particular, the relative intensity of the monoisotopic ion was diminished, and of lower intensity than the first [^{13}C] isotopomer, a relative intensity pattern that is unexpected for a peptide of mass 1031.58 Da, given an empirical formula of $C_{46}H_{78}N_{15}O_{12}$.

The mass isotopomer envelope is consistent with the analyte being a mixture of two peptides, one of monoisotopic *m/z* 1032.58 and a second at a monoisotopic *m/z* of 1033.58. The higher *m/z* peptide could have been a contaminant or it could have been generated from the peptide at *m/z* 1032.58. In the latter case, the most probable explanation for the mass increase was deamidation of the asparagine residue, which, by conversion to an aspartate residue, would increase the mass by 0.985 Da (-NH_2 to -OH). To prove that the atypical profile was a consequence of deamidation, we esterified the peptide mixture to convert carboxyl groups to their methyl esters. The mass shift on esterification would be 14.03 Da. Because the peptide **V_2KVGVNGFGR$_{10}$** would possess a single carboxyl group in the amide form (the alpha carboxyl group), and two in the acid form, esterification should therefore deconvolute the atypical peptide into two products, one esterified at a single position (+14.03 Da), and a second modified in two positions (+28.06 Da). When the peptide mixture was analyzed after esterification, the MALDI-ToF ions in the 1032−1036 *m/z* region disappeared, and two new ions appeared, one representing the single modified amide (*m/z* 1032.58 + 14.03 = 1046.61) and the second reflecting the double modified acid (*m/z* 1033.58 + 28.06 = 1061.64; Figure 2).

From this analysis, it was not possible to assess whether the residue had deamidated *in vivo* or was an artifact of sample preparation and processing. To assess the extent of deamidation of this peptide in the native protein, we treated purified rabbit GAPDH with trypsin and monitored the proteolysis and the partition between the acid and amide variants of the peptide in MALDI-ToF mass spectra (Figure 3; the same experiments were repeated for an in-solution tryptic digest of chicken skeletal muscle soluble proteins and the same behavior was apparent, results not shown). The N-terminal peptide of GAPDH (**VKVGVNGFGR**) was released within a few minutes and was readily detected as the first analyte ion to appear in the MALDI-ToF spectrum. In the early stages of digestion, the mass spectrum of this peptide was entirely consistent with it being exclusively in the amide form. However, as time progressed during proteolysis, the mass spectrum of the peptide showed that the peptide was converted to a mixture of the amide and acid variants, and after 10 h of digestion, the peptide was over 80% in the acid form. The first-order rate constant for this process was approximately 0.0017 min^{-1}, which was higher than the value derived from model peptides; for the sequence NH_2**GVNGG**OH, the first-order rate constant was

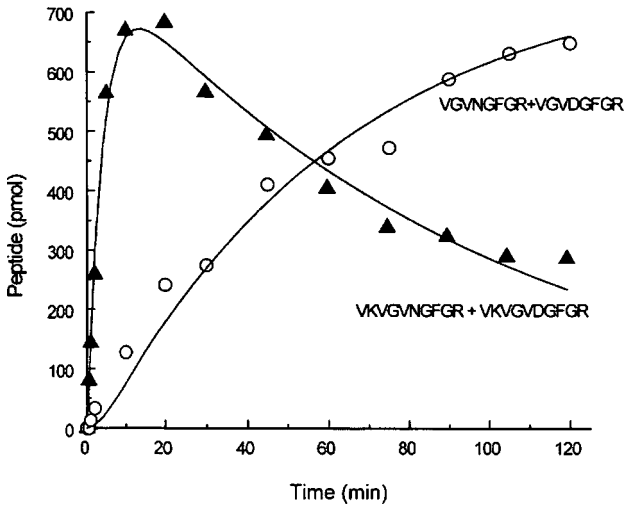


Figure 5. Absolute quantification of proteolysis of the GAPDH N-terminus. Purified rabbit skeletal muscle GAPDH (Sigma, Dorset, U.K.; 1 mg/mL diluted to 0.2 mg/mL with 50 mM ammonium bicarbonate) was digested with trypsin (trypsin/protein 1:10) over 24 h at 37 °C. The N-terminal peptide of GAPDH, of sequence **VKVGVNGFGR** and mass 1041.59 Da, was synthesised by Sigma-Genosys (Dorset, U.K.) and was labelled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10 Da mass offset from the analyte peptide. For quantification of proteolysis, the synthetic peptide was added to digested material in 10% (v/v) formic acid to stop digestion at selected time points. Peptides were analyzed by MALDI-ToF MS, and the relative intensities of analyte peptide and internal standard were used to quantify the amount of peptide released from the protein during incubation with trypsin at 37 °C. Both the N-terminal peptide (**VKVGVNGFGR/VKVGVDGFGR**; *m/z* 1032.59 $[M+H]^+$; closed triangles) and the shorter peptide produced by further proteolysis (**VGVNGFGR/VGVDGFGR**; *m/z* 805.59 $[M+H]^+$; open circles) were monitored. As conversion of asparagine to aspartic acid alters the isotope envelope of the analyte peptide, the composite abundance of the entire isotopic envelope for both analyte and internal standard peptide was summed in each case. The solid lines reflect the fitted curves for the transient appearance of the N-terminal peptide (**VKVGVNGFGR/VKVGVDGFGR**) and the truncated product (**VGVNGFGR/VGVDGFGR**), modelled and fitted as sequential first-order reactions (see text).

previously measured at 0.0004 min^{-1}.[16] However, the buffer conditions for the two experiments are not identical, and pH has a large affect on deamidation rate. The rate of deamidation under these buffer conditions was confirmed using a synthetic peptide of the same sequence; for this peptide, the rate of deamidation was 0.0023 min^{-1}. The higher rate of deamidation of the synthetic peptide might reflect an association between the partially digested protein and the N-terminal peptide which introduced a degree of conformational 'freezing' of the peptide, diminishing the deamidation rate, but this remains conjecture at present.

To investigate the kinetics of both deamidation and proteolysis, a synthetic peptide of sequence VKVGVNGFGR, mass 1041.59 Da, was synthesized and was labeled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10 Da mass offset relative to the natural peptide. This peptide, identical to the N-terminal peptide of GAPDH, was used to monitor the behavior of the peptide, and for quantification.[17] Because the N-terminal peptide itself contains an internal tryptic cleavage site (**VK - VGVDGFGR**), the peptide **VKVGVDGFGR** (summed across acid or amide forms) decreased slowly as digestion continued. We created a model (Figure 4) that took into account the sequential first-order processes of proteolysis (k_1) of the native protein (N_{native}) to release the amide form of the peptide (**VKVGVNGFGR**) followed by deamidation (k_2) to generate the acid form (**VKVGVDGFGR**).

$$N_{native} \xrightarrow{k_1} \text{VKVGVNGFGR} \xrightarrow{k_2} \text{VKVGVDGFGR}$$

Furthermore, the model also included a secondary process of proteolysis of the released peptide in either the acid or amide form to release the ValLys dipeptide. The rate of appearance of the deamidated peptide is given by

$$\text{VKVGV(N/D)GFGR} \xrightarrow{k_3} \text{VK} + \text{VGV(N/D)GFGR}$$

We assumed that the rate of deamidation (k_2) was independent of the N-terminal ValLys dipeptide and that the rate of tryptic removal of the N-terminal dipeptide (k_1) was the same, irrespective of whether the peptide was in acid or amide form. The change in amount (relative to the initial amount of protein, $N_{native(t=0)}$) of the larger peptides (**VKVGVNGFGR** + **VKVGVDGFGR**, $N + D$) as a function of time, is given by

$$N + D = N_{native}\left(\frac{k_1}{k_3 - k_1}(e^{-k_1 t} + e^{-k_3 t})\right) \quad (1)$$

As part of the same process, the shortened peptide (**VGVNGFGR** + **VGVDGFGR**, $N' + D'$) appears according to

$$N' + D' = N_{native}\left(1 - \frac{k_3}{k_3 - k_1}e^{-k_1 t} + \frac{k_3}{k_3 - k_1}e^{-k_3 t}\right) \quad (2)$$

Assuming that the rate of tryptic cleavage is consistent for both acid and amide variants, from these equations, we were able to calculate the second-order rate constants (first-order rate constant divided by protease concentration) for initial release of the large peptide (k_1) and the rate of proteolysis of this large peptide (k_3) (Figure 5). The value of k_1 was estimated to be 1.22 ± 0.025 $min^{-1} \cdot \mu M$ and for k_3, 0.50 ± 0.008 $min^{-1} \cdot \mu M$ (trypsin = 0.2 μM). As expected, the endoproteolytic release of the longer peptide is faster than the release of the N-terminal dipeptide, as trypsin is known to act poorly as a dipeptidyl peptidase. However, the release of the longer peptide is likely to be suppressed by the three-dimensional structure of the protein.

To investigate the effects of the higher order structure of GAPDH on proteolysis and subsequent deamidation, we analyzed the X-ray crystal structure of rabbit GAPDH (PDB code 1J0X.PDB). First, we used the tool NickPred,[18] which although designed to predict sites of proteolytic attack, can generate a comprehensive analysis of the environment of every residue in a protein sequence. The N-terminal region of GAPDH is rather constrained, exhibiting low temperature factors (*B*-values) and low protrusion and accessibility (results not shown). Close inspection of the structure in the vicinity of Asn_6 revealed this region of the polypeptide chain was folded in an extended β configuration, constrained by 14 hydrogen bonds in a network that might be expected to constrain main chain flexibility and therefore reduce the propensity for asparagine deamidation (Figure 6). However, once the peptide was released by proteolysis, deamidation proceeded at a higher rate than that predicted from model studies. These experiments are consistent with the following propositions; that the residue in

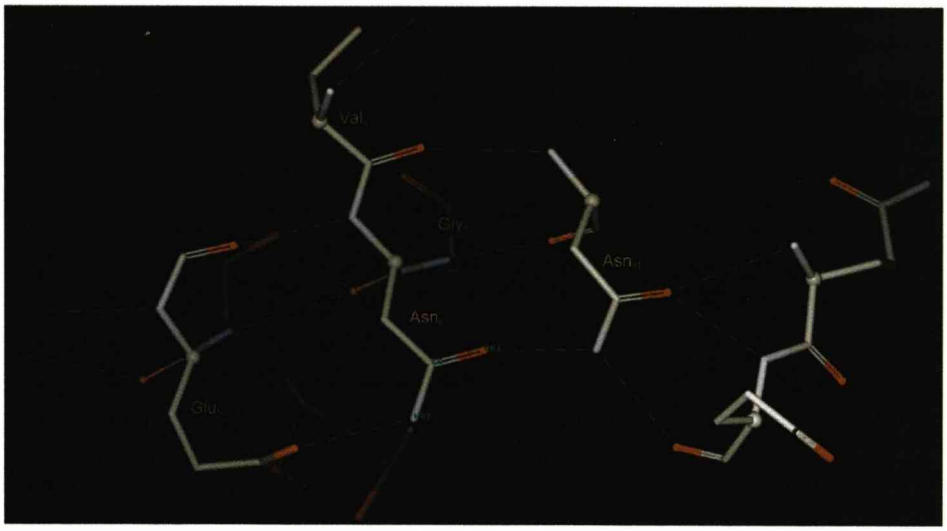


Figure 6. 3D structure of rabbit skeletal muscle GAPDH. X-ray crystal structure of the N-terminal region of rabbit skeletal muscle GAPDH (PDB code 1J0X) highlighting the Asn_6Gly_7 deamidation site and the local hydrogen bonded environment. The green dashed lines denote hydrogen bonds.

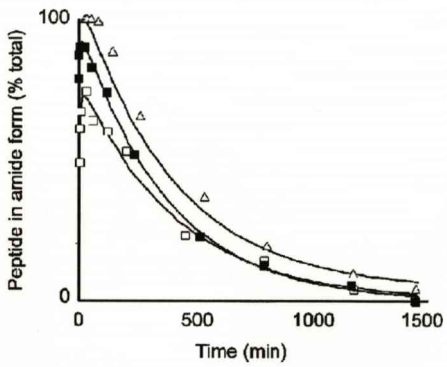


Figure 7. The effect of denaturing protein structure by heating on the rate of deamidation. GAPDH (1 mg/mL diluted to 0.2 mg/mL with 50 mM ammonium bicarbonate) purified from rabbit skeletal muscle (Sigma, Dorset, U.K.) was digested with trypsin in solution at a ratio trypsin/protein 1:100 at 37 °C for 24 h. Prior to digestion, GAPDH was incubated for 1 h at 4 °C (open triangles), 1 h at 60 °C (closed squares), and 1 h at 60 °C followed by 24 h at 37 °C (open squares). For each, deamidation was monitored over 24 h proteolysis and the proportion of acid and amide was calculated from the relative peak intensities of the two ions in MALDI-ToF mass spectra.

the intact protein is exclusively in the amide form, that the tryptic fragment containing the amide residue can undergo deamidation, and that deamidation is not an artifact of the mass spectrometric analysis. Excision of the peptide from the GAPDH structure relieves the constraint in the peptide backbone trajectory, permitting the deamidation reaction to take place. It followed therefore that prior denaturation of the protein might permit deamidation prior to digestion with trypsin. We conducted experiments in which we denatured GAPDH by heating to 60 °C for 1 h before proteolysis (Figure 7), a denaturation treatment that was not sufficient to cause the protein to precipitate. Subsequently, when trypsin was added, the N-terminal peptide was again released rapidly, and the proportion of amide and acid variants of the peptide was assessed as previously described. Under these circumstances, the peptide first released was approximately 80% amide, with a significant proportion of acid form being measurable. This contrasted markedly with proteolysis of the native protein, when the peptide is initially all in the amide form. We attribute this behavior to the increased conformational flexibility of the peptide in the heat-treated protein, such that the peptide could acquire a conformation that allowed deamidation. Further, this unfolded and flexible component might be expected to be hypersensitive to proteolysis and to be released first. As the digestion proceeded, additional peptide in the amide form was released, and the proportion of amide therefore increased transiently, until the deamidation reaction dominated the peptide profile. When the functions derived previously were used, we obtained a value for deamidation of 0.0023 min^{-1}, in close agreement with that observed previously. If the heat-treated peptide was allowed to incubate at 37 °C for 24 h after the 60 min denaturation period at 60 °C, and then proteolyzed with trypsin, the peptide first released was now only 50% in the amide form, consistent with extensive deamidation prior to proteolysis, consequential to denaturation. Again, as expected, proteolysis led to the slower release of peptide that was constrained and unable to deamidate, and there was a transient increase in the proportion of amide which again decayed at the same rate as observed previously ($k_2 = 0.0024$ min^{-1}). The behavior of the system was consistent with the GAPDH preparation being 76% in the amide form, and 26% in a denatured form that was then rapidly proteolyzed to generate the free acid form of the peptide. The effect of denaturation on the availability of the N-terminal peptide of GAPDH for deamidation is quite striking and defines the importance of monitoring the two processes of proteolysis and deamidation simultaneously, especially as this effect is only observed upon loss of higher order structure, and not upon increasing concentration of protease (results not shown).

Conclusions

Deamidation is recognized as a potential source of microheterogeneity in protein structure, and it may play a significant role as a biological 'timer' that is mediated nonenzymatically.[1-5] Although many studies have emphasized the deamidation of short, flexible peptides, protein deamidation can be limited by higher order structure and might only occur at the peptide level

following proteolytic release.[8] The ease with which some peptides deamidated could then lead to the erroneous interpretation of a deamidation event as occurring in the intact protein. Difficulties of measuring deamidation have been discussed,[19] and analyses often use electrospray ionization mass spectrometry[6] and reversed-phase chromatographic matrices[8] to resolve acid and amide variants of a peptide, precluding analysis of complex mixtures. There is also considerable scope for MALDI sample ionization, which, when coupled with a simple esterification reaction, can clearly identify and characterize such deamidation variants. We suggest that there may be merit in closer examination of the isotope distribution profile of peptide mass fingerprints, to search for anomalies such as that noticed here. In particular, it is advantageous to monitor deamidation and proteolysis simultaneously when characterizing post-translational behavior of known proteins and peptides. This will also unravel information about the higher order structure of a protein, the influence of which not only on proteolysis but also on subsequent modifications to newly accessible regions of the protein, is paramount.

Acknowledgment. We are grateful to Dr. Gary Evans, Genus plc, for his interest in this work. This work has been supported by the BBSRC, EPSRC (EP/D013623) and Genus plc. We are grateful to Dr. D. H. Robertson for instrumentation support.

References

(1) Robinson, A. B.; Rudd, C. J. Deamidation of glutaminyl and asparaginyl residues in peptides and proteins. *Curr. Top. Cell. Regul.* **1974**, *8* (0), 247–295.

(2) Geiger, T.; Clarke, S. Deamidation, isomerization, and racemization at asparaginyl and aspartyl residues in peptides. Succinimide-linked reactions that contribute to protein degradation. *J. Biol. Chem.* **1987**, *262* (2), 785–794.

(3) Friedman, A. R.; Ichhpurani, A. K.; Brown, D. M.; Hillman, R. M.; Krabill, L. F.; Martin, R. A.; Zurcher-Neely, H. A.; Guido, D. M. Degradation of growth hormone releasing factor analogs in neutral aqueous solution is related to deamidation of asparagine residues. Replacement of asparagine residues by serine stabilizes. *Int. J. Pept. Protein Res.* **1991**, *37* (1), 14–20.

(4) Deverman, B. E.; Cook, B. L.; Manson, S. R.; Niederhoff, R. A.; Langer, E. M.; Rosova, I.; Kulans, L. A.; Fu, X.; Weinberg, J. S.; Heinecke, J. W.; Roth, K. A.; Weintraub, S. J. Bcl-xL deamidation is a critical switch in the regulation of the response to DNA damage. [erratum: *Cell* **2003**, *115* (4), 503] *Cell* **2002**, *111* (1), 51–62.

(5) Weintraub, S. J.; Manson, S. R. Asparagine deamidation: a regulatory hourglass. *Mech. Ageing Dev.* **2004**, *125* (4), 255–257.

(6) Robinson, N. E.; Robinson, A. B.; Merrifield, R. B. Mass spectrometric evaluation of synthetic peptides as primary structure models for peptide and protein deamidation. *J. Pept. Res.* **2001**, *57* (6), 483–493.

(7) Doherty, M. K.; McLean, L.; Hayter, J. R.; Pratt, J. M.; Robertson, D. H.; El-Shafei, A.; Gaskell, S. J.; Beynon, R. J. The proteome of chicken skeletal muscle: changes in soluble protein expression during growth in a layer strain. *Proteomics* **2004**, *4* (7), 2082–2093.

(8) Krokhin, O. V.; Antonovici, M.; Ens, W.; Wilkins, J. A.; Standing, K. G. Deamidation of -Asn-Gly- sequeneces during sample preparation for proteomics: consequences for MALDI and HPLC-MALDI analysis. *Anal. Chem.* **2006**, *78*, 6645–6650.

(9) Stroop, S. D. A modified peptide mapping strategy for quantifying site-specific deamidation by electrospray time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2007**, *21*, 830–836.

(10) Jedrzejewski, P. T.; Girod, A.; Tholey, A.; Konig, N.; Thullner, S.; Kinzel, V.; Bossemeyer, D. A conserved deamidation site at Asn 2 in the catalytic subunit of mammalian cAMP-dependent protein kinase detected by capillary LC-MS and tandem mass spectrometry. *Protein Sci.* **1998**, *7* (2), 457–469.

(11) Nilsson, M. R.; Driscoll, M.; Raleigh, D. P. Low levels of asparagine deamidation can have a dramatic effect on aggregation of amyloidogenic peptides: implications for the study of amyloid formation. *Protein Sci.* **2002**, *11* (2), 342–349.

(12) Cournoyer, J. J.; Lin, C.; O'Connor, P. B. Detecting deamidation products in proteins by electron capture dissociation. *Anal. Chem.* **2006**, *78*, 1264–1271.

(13) Bogdan, I.; Coca, D.; Rivers, J.; Beynon, R. J. Hardware acceleration of processing of mass spectrometic data for proteomics. *Bioinformatics* **2007**, *23* (6), 724–731.

(14) Pratt, J. M.; Simpson, D. M.; Doherty, M. K.; Rivers, J.; Gaskell, S. J.; Beynon, R. J. Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat. Protoc.* **2006**, *1* (2), 1029–1043.

(15) Rivers, J.; Simpson, D. M.; Robertson, D. H. L.; Gaskell, S. J.; Beynon, R. J. Absolute multiplexed quantitative analysis of protein expression during muscle development using QconCAT. *Mol. Cell. Proteomics* **2007**, *6*, 1416–1427.

(16) Robinson, N. E.; Robinson, Z. W.; Robinson, B. R.; Robinson, A. L.; Robinson, J. A.; Robinson, M. L.; Robinson, A. B. Structure-dependent nonenzymatic deamidation of glutaminyl and asparaginyl pentapeptides. *J. Pept. Res.* **2004**, *63* (5), 426–436.

(17) Kirkpatrick, D. S.; Gerber, S. A.; Gygi, S. P. The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods (Duluth)* **2005**, *35* (3), 265–273.

(18) Hubbard, S. J. The structural aspects of limited proteolysis of native proteins. *Biochim. Biophys. Acta* **1998**, *1382* (2), 191–206.

(19) Lindner, H.; Helliger, W. Age-dependent deamidation of asparagine residues in proteins. *Exp. Gerontol.* **2001**, *36* (9), 1551–1563.

PR070425L